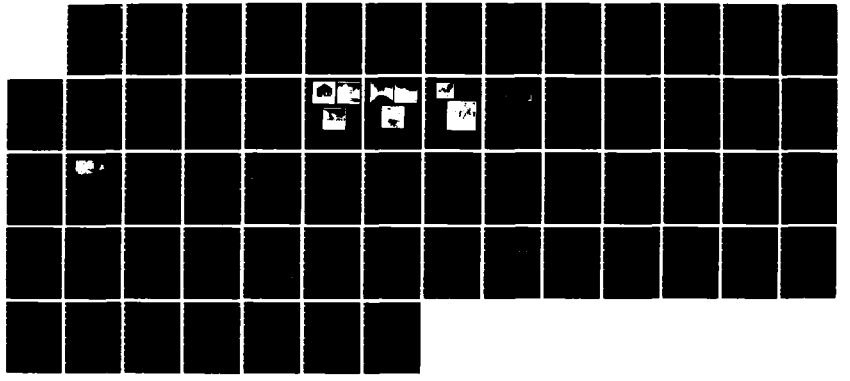
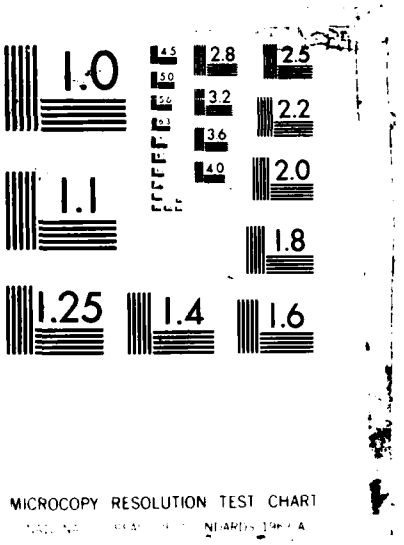
AD-A179 116 REPRESENTATION AND CONTROL IN THE INTERPRETATION OF COMPLEX SCENES(U) MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER AND INFORMATION S A R HANSON ET AL 1987 1/1

UNCLASSIFIED AFOSR-TR-87-0301 F49620-83-C-0099 F/G 9/2 NL





Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-0301
6a. NAME OF PERFORMING ORGANIZATION: University of Massachusetts; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NM
6c. ADDRESS (City, State and ZIP Code): Dept. of Computer Science, Amherst, MA 01003	7b. ADDRESS (City, State and ZIP Code): Bldg 410, Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable): NM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F49620-83-C-0099
8c. ADDRESS (City, State and ZIP Code): Bldg 410, Bolling AFB DC 20332-6448	10. SOURCE OF FUNDING NOS. — PROGRAM ELEMENT NO.: 61103F; PROJECT NO.: 2304; TASK NO.: A7; WORK UNIT NO.:
11. TITLE (Include Security Classification): Knowledge-based control in the interpretation of Complex Scenes	
12. PERSONAL AUTHOR(S): Professor Hanson	
13a. TYPE OF REPORT: Final; 13b. TIME COVERED: FROM 83Apr.01 TO 85Mar31; 14. DATE OF REPORT (Yr., Mo., Day):	15. PAGE COUNT
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB GR	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Let us summarize key characterisitics of our design that are relevant to the issues involved in the construction of a general interpretation system; the VISIONS system incorporates:

- effective (although imperfect) segmentation processes that are knowledge-independent and parameterized to control the degree of sensitivity in the output;
- an intermediate symbolic representation to serve as an interface between sensory data and knowledge;

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL: Brian W. Woodruff	22b. TELEPHONE NUMBER (Include Area Code): (202) 767-5026 — 22c. OFFICE SYMBOL: NM

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

9003f

Representation and Control in The Interpretation of Complex Scenes

Final Scientific Report

for the period

October 1, 1984 to September 30, 1985

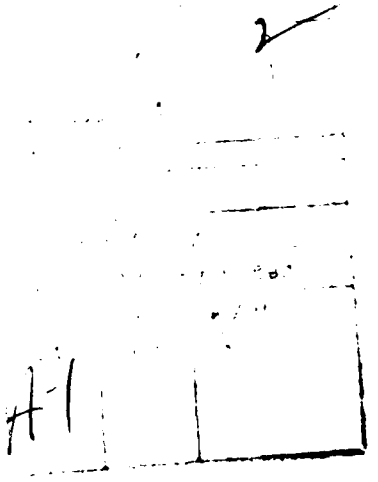
Submitted by

Allen R. Hanson, Principal Investigator
Edward M. Riseman, Co-Principal Investigator

Grant Number AFOSR-85-0005

(continuation of contract F49620-83-C-0099)

Contents



1 INTRODUCTION

The system we have been developing, called VISIONS, is an investigation into issues of general computer vision. The goal is to provide an analysis of color images of outdoor scenes, from segmentation through symbolic interpretation. The output of the system is intended to be a symbolic representation of the three-dimensional world depicted in the two-dimensional image, including the naming of objects, their placement in three-dimensional space, and the ability to predict from this representation the rough appearance of the scene from other points of view.

The emphasis of the research over the past year has been on three issues critical to furthering our understanding of machine vision. The first area addresses the issue of image segmentation and the failure of recent research to provide robust procedures applicable to complex imagery. We have begun to develop an expert segmentation system which would combine existing techniques using image specific knowledge to improve segmentation results. Two general segmentation approaches, based on regions and lines, have been developed towards this end. The second area focusses on the use of domain knowledge in the interpretation task. Domain knowledge is embedded in high-level structures called schemas; knowledge-based expectations from the schema are used to focus the interpretation process. Terry Weymouth's forthcoming Ph.D. Thesis explores this issue in detail. The third area focusses on techniques for controlling the use of system resources during interpretation and on ways of resolving conflicting partial interpretations. Leonard Wesley's forthcoming Ph.D. Thesis demonstrates how the Shafer-Dempster formalism for evidential reasoning may be extended to a control mechanism for interpretation.

The remainder of this report is continued in four sections. Section 2 provides a very brief overview of the VISIONS system. Section 3 discusses image segmentation, Section 4 explores image interpretation, and Section 5 presents the preliminary design of a vision system based on the use of evidential reasoning as a control mechanism.

2 Overview of the VISIONS System Approach

In response to the issues that we have outlined we have evolved over the last twelve years a general partially working image understanding system. Our system design embodies certain assumptions about the process of transforming visual information. The first assumption underlying our work is that the initial computation proceeds in a bottom-up fashion by extracting information from the image without knowledge of its contents. A second basic assumption is that every stage of processing is inherently unreliable. A third assumption is that local ambiguity and uncertainty in object hypotheses to a great extent can only be removed by satisfying expected relations between scene and object parts that are stored in a knowledge base about the domain.

Thus, the successful functioning of our system will involve extracting image events which are then used to hypothesize scene and object parts for quick access to knowledge structures, called schemas, embodying the object descriptions and contextual constraints from prototypical scene situations. The hierarchically organized schemas embody interpretation strategies for top-down control of intermediate grouping strategies and allow feedback from high-level hypotheses to low-level processing.

The general strategy by which the VISIONS system operates is to build an intermediate symbolic representation (ISR) of the image data using segmentation processes which initially do not make use of any knowledge of specific objects in the domain. From the intermediate level data, a partial interpretation is constructed by associating an object label with selected groups of the intermediate tokens. The object labels are used to activate portions of the knowledge network related to the hypothesized object. Once activated, the procedural components of the knowledge network direct further grouping, splitting and labelling processes at the intermediate level to construct aggregated and refined intermediate events which are in closer agreement with the stored s mbolic object description. Figure 2.1 is an abstraction of the multiple levels of representation and processing in the VISIONS system. Communication between these levels is by no means unidirectional; in

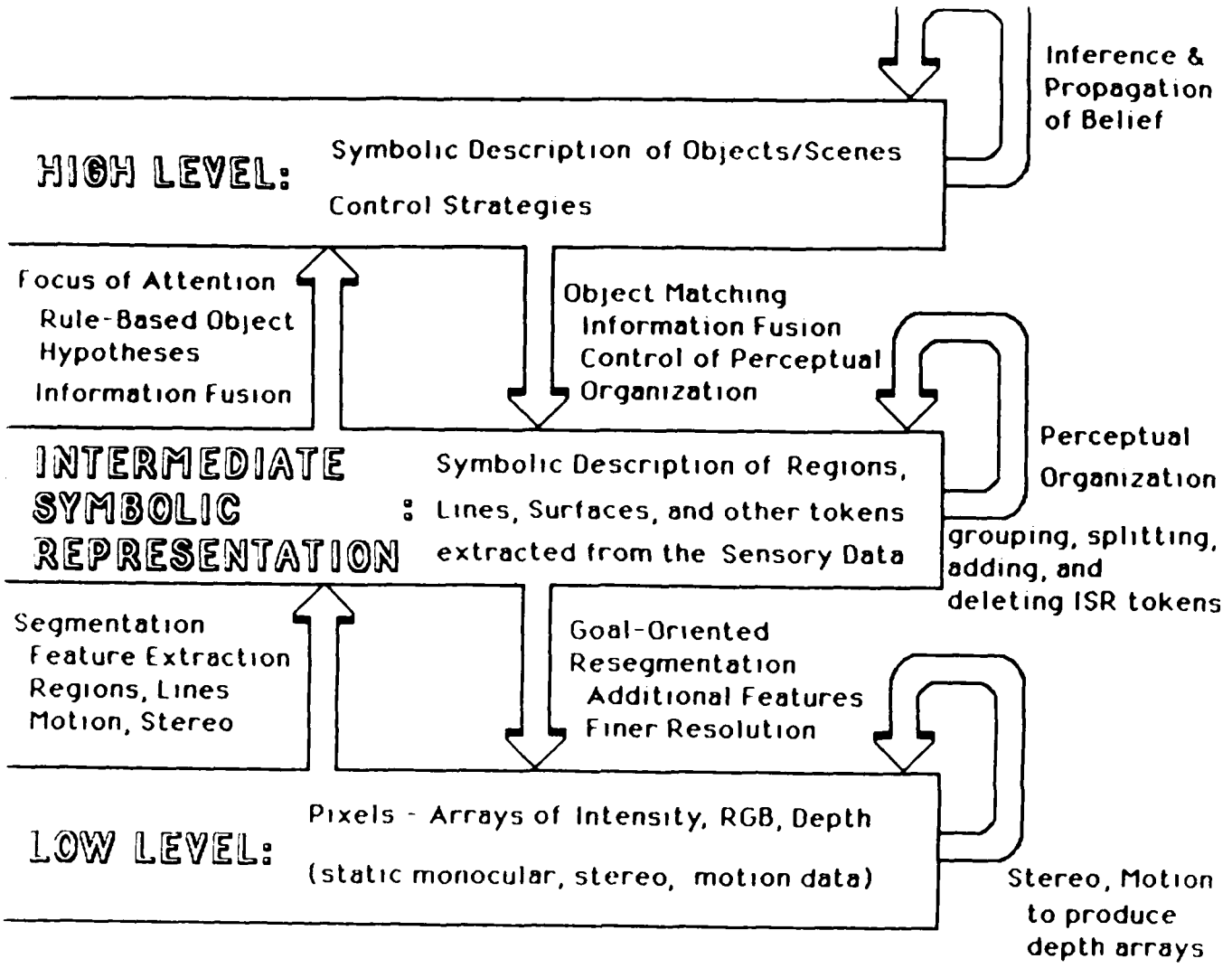


Figure 2.1

Image Interpretation

Multiple levels of representation and processing in VISIONS.

most cases, recognition of an object or part of a scene at the high level establishes a strategy for further manipulating the intermediate level primitives within the context provided by the partial interpretation, and for feedback for goal-directed resegmentation. Although the following discussion is based primarily on 2D abstractions of the image data (such as regions and lines), it should be clear that the general ideas extend naturally to 3D abstractions such as surfaces as well as to attributes such as motion and depth.

Let us briefly describe the three levels of representation:

1. Low-Level - Here arrays of direct numerical sensory data are stored, including the results of algorithms which produce point/pixel data in register with the sensory data (e.g. a depth map produced from stereo point matching).

2. Intermediate-Level - This is referred to as the Intermediate Symbolic Representation (ISR) because symbolic tokens for regions, lines, and surfaces with attribute lists are constructed for the image events that have been extracted from the low-level data. Aggregrate structures produced by grouping and/or extending the primitive image events or other aggregates are also reprsented symbolically as tokens in the ISR.

3. High-Level - The knowledge base (called Long Term Memory or LTM) consists of a semantic network of schema nodes, each of which has a declarative and procedural component. The network is organized in terms of a compositional hierarchy of PART-OF relations and a subclass hierarchy of IS-A relations.

Now let us consider some of the stages of processing in a bit more detail.

1. Motion and Stereo

   - Multiple frames can be analyzed via motion and stereo algorithms to produce displacement fields and from these displacement fields depth arrays can be extracted. Note that if line correspondences are determined, then the result can be viewed as a direct

extraction of a symbolic token at the intermediate symbolic representation and a form of low-level segmentation process.

2. Segmentation

- Segmentation processes are applied to the sensory data to form a symbolic representation of regions and lines and their attributes such as color, texture, location, size, shape, orientation, length, etc. The region and line representations are integrated so that spatially related entities in either can be easily accessed. If depth arrays are available, then 3D surfaces and their attributes can be extracted. Two-dimensional motion attributes can also be associated with regions and lines, while 3D motion attributes can be associated with surfaces and their boundaries.

3. Focus-of-Attention

- Object hypothesis rules are applied to the region, line, and surface tokens in the ISR to rank-order candidate object hypotheses [1,28,33,34]. This initial mapping of spatially organized tokens to object labels provides an effective focus-of-attention mechanism to initiate semantic processing. The rules can also be viewed as sets of constraints on the range of attribute values on the ISR tokens; e.g. constraints on the color of region tokens, or length of line tokens, or orientation of surface tokens. Relational rules applied to tokens of a different type (e.g. line to region relations) allow fusion of information across multiple representations.

4. Object Matching

- More complex object-dependent interpretation strategies are represented in a procedural form within the schema knowledge structures [8,25]. These local control strategies provide top-down control over the interpretation process. Partial interpretations are extended from "islands of reliability" as in the HEARSAY paradigm [7,16]. General

knowledge of objects and scenes is organized around a network of standard 2D views of 3D objects. In cases of simple 3D shapes, such as the planar surfaces forming a "house" volume, 3D models and associated processing strategies are employed. Verification strategies exploit particular spatial relationships between the hypothesized object and other specific expected object labels or image features.

5. Perceptual Organization

- Intermediate grouping algorithms, currently being developed, e.g., reorganize the region and line elements into larger aggregrate structures which are expected to more closely match expected object structures. These include line grouping and region merging capabilities and are intended to be applied in a bottom-up or top-down manner. There are significant advantages in selective application of knowledge-directed perceptual organization mechanisms, and the schemas provide an effective means for specification of the relevant control knowledge. We hope to evolve similar intermediate grouping strategies for complex 3D shape representations in the future.

6. Goal-Oriented Resegmentation

- Feedback to the lower-level processes for more detailed segmentation can be requested in cases when interpretation of an area fails, when an expected image feature is not found, or when conflicting interpretations occur. It may also be of utility in motion sequences for extraction of particular objects that have been found in previous frames. Both the region and line algorithms have parameters for varying the sensitivity and amount of detail in their segmentation output. The development and control of such strategies, as well as the integration of their results is currently under examination in an effort to build a low-level executive.

7. Inference

   - Due to the inherent ambiguities in both the raw image data and the extracted intermediate representations, object hypotheses will have a high degree of uncertainty. In order to combine this information into a coherent view of the world, two capabilities are being developed: the capability to combine uncertain evidence from multiple sources of knowledge [27], and the ability to propagate confidences (of beliefs or probabilities) through the network of schema nodes [17,31]. The latter capability will allow inferential capabilities about the presence of object/scene parts for control of the partially completed interpretation. Both heuristic approaches and theoretical formulations similar to the Dempster-Shafer theory of evidential reasoning [6,29] are currently being explored.

2.1 Summary of Design Principles

Let us summarize key characterisitics of our design that are relevant to the issues involved in the construction of a general interpretation system; the VISIONS system incorporates:

- effective (although imperfect) segmentation processes that are knowledge-independent and parameterized to control the degree of sensitivity in the output;
- an intermediate symbolic representation to serve as an interface between sensory data and knowledge;
- a focus-of-attention mechanism for initial object hypotheses to deal with uncertainty and unreliability in the early stages of the interpretation process when there is a lack of context;
- procedural knowledge encoded as object specific interpretation strategies within schemas;
- knowledge-directed grouping of the ambiguous, incomplete, and uncertain perceptual events extracted from the sensory data;

- fusion of the multiple sensory data, intermediate representations, and multiple sources of knowledge during the interepretation process;
- knowledge-directed feedback to lower level vision mechanisms to extract additional information from the sensory data.

Most of our research, particularly that funded by the AFOSR, is oriented towards incremental expansion and extensions to the existing system.

3 Segmentation Algorithms

Probably the most fundamental problem blocking knowledge-based vision development has been the lack of stable low-level vision algorithms that can produce a reasonably useful intermediate representation. Vision systems must deal with the problem of dynamically transforming the massive amounts of sensory data (in an image of reasonable resolution there are $512 \times 512 \cong 1/4$ million pixels) into a much smaller set of image events, such as regions of homogeneous color and texture, straight lines, and local surface patches, to which propositions will be attached.

There is little doubt that the segmentation problem[1] is a very difficult and ill-formed problem [9]. There is no ideal or "correct" segmentation because that is a function of the goals of the interpretation system. The 2D appearance of objects and their parts is affected by variations in lighting, perspective distortion, varying point of view, occlusion, highlights, and shadows. In addition objects and their parts may or may not be visually distinguished depending upon their color and the background. Thus, one cannot predict what initial image information can be extracted that is relevant to the recognition of objects. The only thing one should count on is that the process is unreliable; some useful information can be extracted while many errors will also occur, and that there is probably no optimal solution in any non-trivial image domain. In general there is no universally acceptable set of parameter settings for any algorithm which are guaranteed to extract the desired image events without also generating additional non-optimal or undesirable events. For a given parameter setting, a region segmentation might be too fragmented in one area of an image (i.e. too many regions) while being overmerged in another area of the image (i.e. too few regions). As parameters are varied the partitioning will change, but never will a result be produced that is optimal or near-optimal throughout the scene. The same is true of line and surface extraction algorithms; they will produce fragmented and overmerged events in a varied and unpredictable

[1] Note that we sometimes will use the term segmentation loosely to cover not only the usual definition of partitioning an image into connected, non-overlapping sets of pixels, but also other low-level processes such as line extraction.

manner.

3.1 Segmentation vs. the Recovery of 3D Surfaces

As we have just noted, for the past decade there has been major controversy in the field of computer vision concerning the meaningfulness and validity of the segmentation problem. Some researchers seem to restrict their criticisms to the actual partitioning of images via region segmentation, but feel more open to the process of line extraction. It is our strong opinion that while both suffer from similar types of unreliability, both result in descriptions which contain useful information and therefore should not be distinguished with respect to the issues being discussed. Researchers who are more theoretically inclined, and whose work is sometimes referred to as computational vision [3,12,13,18,30], have concluded that the recovery of information about the 3D surfaces in the visible environment is the most (and some have implied only) appropriate problem. In the extreme this leads to systems which are restricted to using only those processes which directly recover 3D information (stereo, motion, shape from shading, shape from texture, and in general shape from x, where x is a monocular source of information).

It is our position that even when we have the best intermediate level surface information possible, the complexity of the natural world will leave us facing many of the difficult issues that we have been discussing. Consider the problem of interpreting a natural environment such as a typical crowded city street scene even if one had a perfect depth map for all the visible surfaces; i.e., in addition to the original color information at each pixel, we are assuming that the distance to the corresponding visible surface element at each pixel would also be available. How should one partition the information into meaningful entities such as surfaces, parts of objects, and objects? And then how could this be interfaced to the knowledge base so that control of the interpretation process is feasible? Given that many initial local hypotheses will be inherently uncertain and unreliable, how do we achieve globally consistent and reliable integration of the information?

This, in fact, is exactly the set of problems that we face with 2D data of regions and lines. One

cannot escape the problem of 3D segmentation, and in some manner must face the partitioning of the depth map into surface patches of various types and the extraction of 3D lines via depth and orientation discontinuties. We believe that the principles and approaches presented here for analysis of 2D color data of static monocular images will also be applicable to the processing of 3D depth data derived from stereo and laser ranging devices, as well as 2D and 3D motion data derived from a sequence of images.

3.2 Segmentation Algorithms

There are a variety of sources of low-level information available in an image of a scene which can be used to form an initial description of the structure of the image. The extraction of this description, in terms of significant image "events" is an important precursor to the construction of a more abstract description at the semantic level. It is unlikely that any single descriptive process will produce a description which is adequate for an unambiguous interpretation of the image. Rather, multiple processes are required, each of which views the image data in a different way and each of which produces a partial description which may be incomplete or errorful. As we shall see in later sections, we view one of the functions of the perceptual grouping processes to be the resolution of the multiple descriptions into consistent higher-level symbolic descriptions of the image. Consequently, for the entire history of our project we have continued the development of multiple low-level algorithms which are reasonably robust across a variety of task domains, while recognizing that no single algorithm is sufficient.

We will sometimes refer to the entire set of low-level processes for extracting image events as segmentation algorithms, including region and line extraction algorithms, even though some of them do not actually partition in the image into disjoint subsets; thus, sometimes we loosely use "low-level processes" and "segmentation processes" interchangeably.

It should also be noted that we view all of these algorithms operating initially in a bottom-up mode without any use of knowledge of the task domain. However, we allow rough "sensitivity"

settings on parameters of the algorithm, such as "low", "medium", and "high", to be set a priori in order to extract a larger or smaller number of intermediate tokens; these parameters may also be set or modified by other components of the system and the application of the process may be localized to a specified subimage.

While a variety of low-level algorithms have been developed, there are two primary low-level algorithms that are currently in use in our environment a) a region segmentation algorithm that is based upon analysis of histogram peaks and valleys in local subimages; and b) a straight line extraction algorithm that segments the intensity surface into connected subsets of pixels with similar gradient orientation.

These algorithms, in addition to a region merging segmentation algorithm, are being integrated into a low-level executive for directing goal-oriented low-level procesing via feedback from interpretation processes [14]. These algorithms are discussed below.

3.2.1 Histogram-Based Region Segmentation

The region segmentation technique that we employ was first developed by Nagin [21] and later extended by Kohler [15]. The algorithm is currently being further extended by R. Beveridge. The approach is in the spirit of the Ohlander-Price algorithms [23,26]. However, since both their algorithms should be viewed as instances of histogram-based region segmentation, we believe that generally our approach will be more robust and computationally efficient because of the problems of recursive decomposition of global histograms in the Ohlander-Price approach [10]. Our algorithm allows a fixed number of windows to be processed on one pass and definitely avoids the problem of hidden clusters in global histograms in a recursive decomposition approach.

The histogram-based region segmentation algorithm involves detecting clusters in a feature histogram, associating labels with the clusters, mapping the labels onto the image pixels, and then forming regions of connected pixels with the same label. The process of global histogram labeling causes many errors to occur because global information will not accurately reflect local image events

that do not involve large numbers of pixels, but which nevertheless are quite clear. Much of the focus of the algorithms will be to organize the segmentation process around local histograms from local windows and then have a postprocesing stage merge regions that have been arbitrarily split along the artificial window boundaries.

The Nagin algorithm overcomes this difficulty by partitioning the image into NxN subimages (usually N = 16 or 32) called sectors; the histogram segmentation algorithm is applied independently to each sector. Thus, each sector receives the full focus of the cluster detection process and many of the problems of cluster overlap and "hidden" clusters are significantly reduced. The partitioning of the image into sectors has an obvious weakness. If an adjacent sector has a visually distinct region which does not overlap the central sector sufficiently, it is quite possible that the cluster will be undetected in the central sector. The small region representing the intrusion into the central sector will then be lost. Nagin attempted to limit this problem, while still preserving the locality of the histogram, by expanding each sector to overlap adjacent sectors by 25%.

The Kohler algorithm improves the clustering step by determining correspondences between peaks in adjacent sectors and by adding candidate peaks from surrounding sectors to the set of peaks selected for the central sector. Thus, small variations between peak labels will be accounted for and if the 25% overlap is not sufficient, small intrusions of regions from one sector to the next will more likely result in a peak to be added. The augmented set of peaks forms the basis of the labeling process.

The basis for mapping labels to each pixel in the image is a histogram cluster analysis. While we will present our simple cluster extraction algorithm via histograms (i.e. the use of only a single feature), if a 2D clustering algorithm for associating pixel labels with two feature values is available, then the rest of the algorithm remains the same. Theoretically, any N-dimensional clustering algorithm could be used, but one must keep in mind that computational considerations are primary since this clustering algorithm will be executed many times across the image and is

only one small step in the entire interpretation process. The one-dimensional cluster extraction algorithm determines local maxima and local minima in the histogram, and then sequentially chooses the next 'best' maxima, which is that maxima for which the ratio of the local maximum to its adjacent valley is largest and the distance to any previously selected local maximum is greatest. The net result is a set of local maxima and their relationships to the valley separating them, each of which can be considered as a distinct cluster and given a unique label. Each of these labels are then mapped back to the subimage and a connected components is used to generate a set of labeled regions.

The artificial sector boundaries are removed by a region merging algorithm which considers adjacent region across sector boundaries. The merge/no merge decision mechanism compares the global mean and variance of the two regions, as well as the local mean and variance of the regions near the sector boundary. This remerging algorithm is being generalized to use a set of modular feature statistics so that many different remerging strategies can be used and so it can be applied separately from the histogram-based region segmentation algorithm under discussion here.

The algorithm can be summarized in five steps:

1. Subdivide the image into sectors and select cluster labels in each expanded sector (Figure 3.1).

2. Analyze cluster labels from adjacent sectors for augmenting the label set.

3. Segment each sector using the expanded set of labels by mapping each pixel value to a label and applying a connected components algorithm.

4. Remove sector boundaries by merging similar regions.

5. Perform small region suppression, which often reduces the number of regions by a factor of 4 (Figure 3.2).

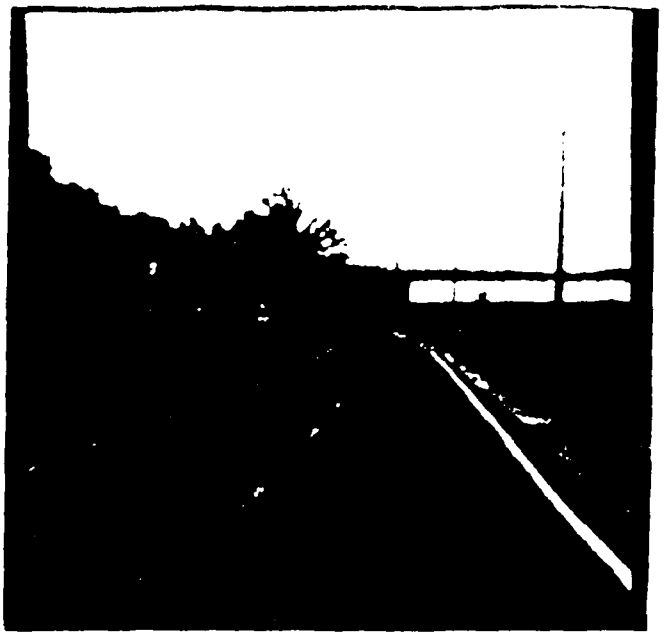
(a)

(b)

(c)

Original images. These images are representative samples from a larger data base. All are digitised to 512 × 512 spatial resolution, with 8 bits in the red, green, and blue components.

(d)

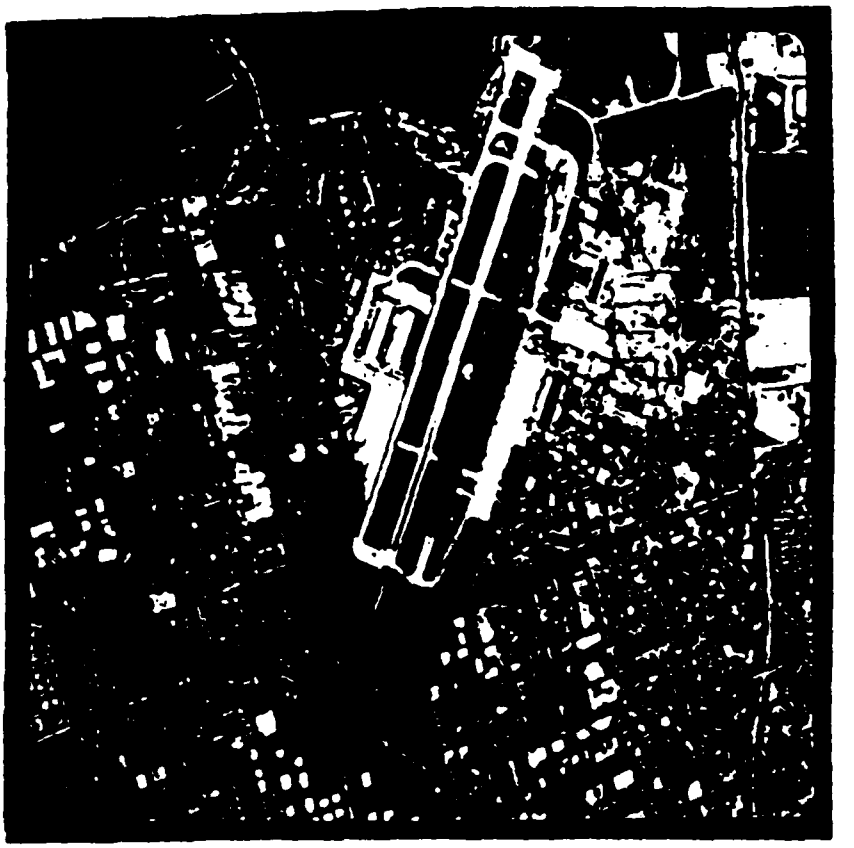


(e)

(f)

(Continued)

(g)



(h)

Figure 3.C, continued

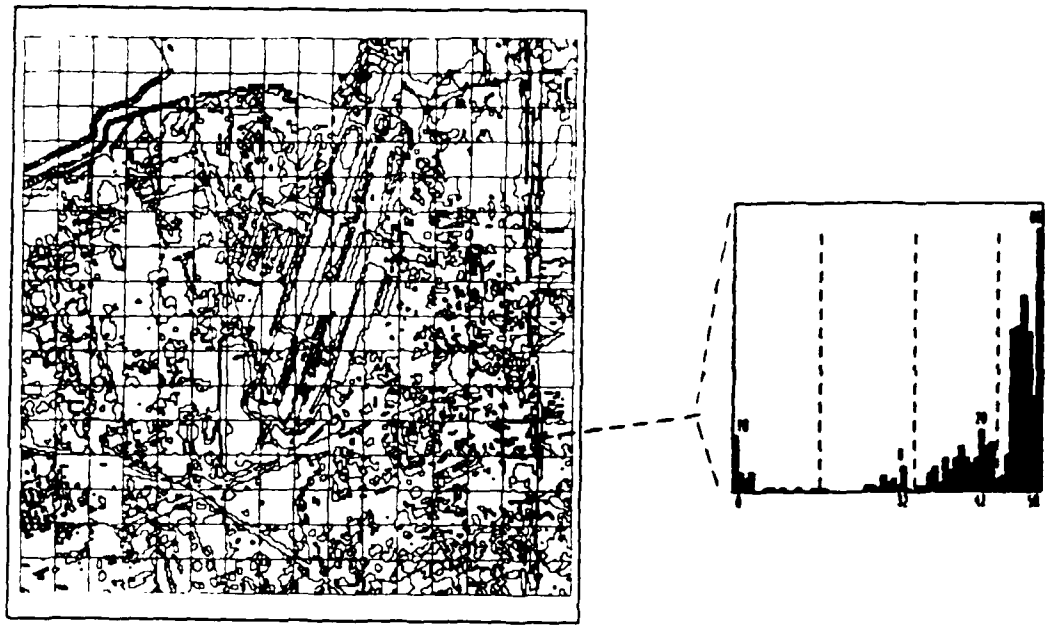


Figure 3.1.

Segmentation of each sector using the expanded set of labels.

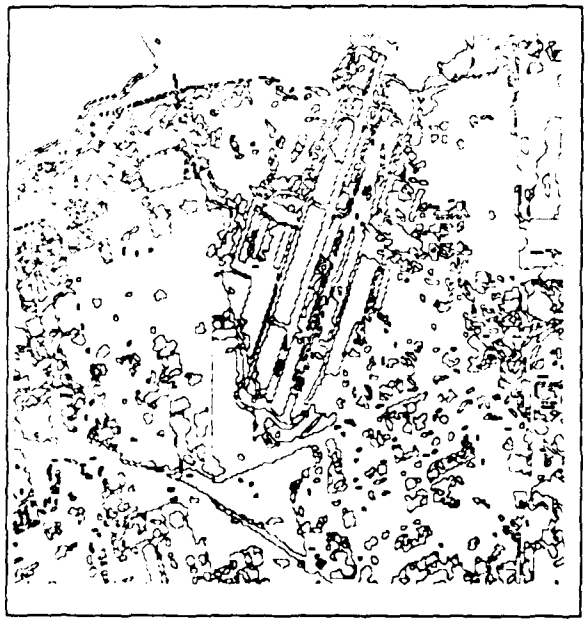


Figure 3.2.

Segmentation after removing sector boundaries and elimination of small regions

Figure 3.3 shows additional segmentations using this algorithm on some of the images in Figure 3.0.

3.2.2 Extracting Straight Lines and Line-Based Texture Features

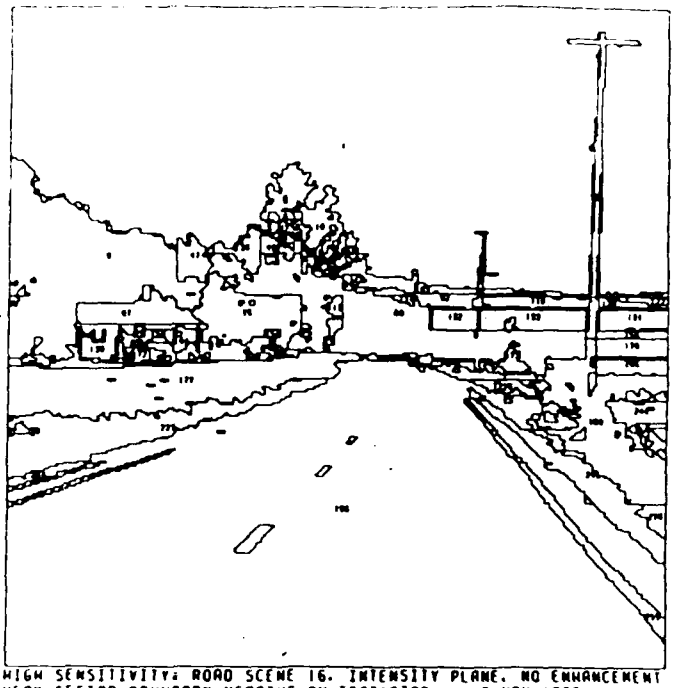
Despite the large amount of research appearing in the literature, effective extraction of linear boundaries has remained a difficult problem in many image domains. The technique presented here [4] was motivated by a need for a straight line extraction method which can find straight lines, possibly long and possibly of very low contrast, in reasonably complex images.

A key characteristic of our approach that distinguishes it from most previous work is the global organization of the intensity surface into a "supporting edge context" prior to making any decisions about the relevance of local intensity changes. Pixels are grouped into edge support regions of similar gradient orientation, and then the associated intensity surface is used to determine a gradient magnitude weighted planar fit from which a representative line is extracted. A set of line attributes is extracted from the line-support region, the weighted planar fit, and the representative line. These attributes can then be filtered for a variety of purposes.

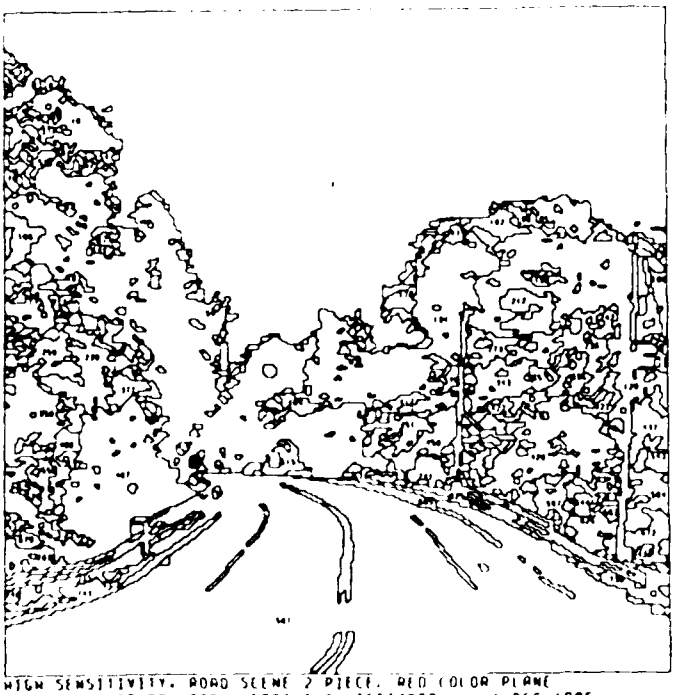
Figure 3.4(a) is a surface plot of a 32x32 intensity image. The vector field drawn in figure 3.4(b) shows the corresponding gradient image (where the length of the vector encodes gradient magnitude) computed using two 2x2 edge masks.

An extremely simple and computationally efficient process is employed to group the local gradients into regions on the basis of the orientation estimates. The 360 degree range of gradient directions is arbitrarily partitioned into a small set of regular intervals, say eight 45 degree partitions or sixteen 22.5 degree partitions. Pixels participating in the edge-support context of a straight line will tend to be in the same edge orientation partitions and adjacent pixels that are not part of a straight line will tend to have different orientations. A simple connected components algorithm can be used to form distinct region labels for groups of adjacent pixels with the same orientation label (Figure 3.4(c)). The great degree of fragmentation into many small regions of very low gra-

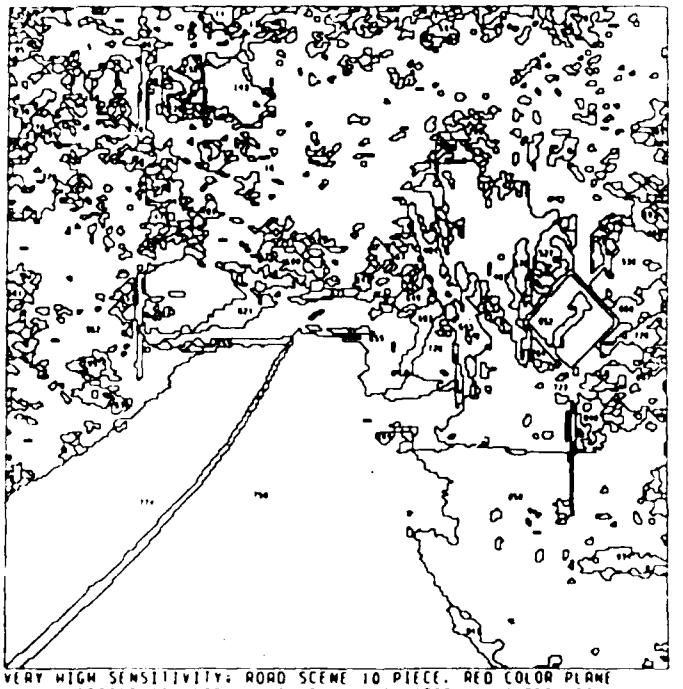
HIGH SENSITIVITY: HOUSE IMAGE 1. RED COLOR PLANE
NORMAL SECTOR BOUNDARY MERGING ON IM1RB. 7-NOV-1985
SECTOR SIZE = 16. WEYMOUTH ENHANCEMENT. DATA RANGE 1.0 TO 61.0
MIN-PEAK-DISTANCE 7 PERCENT. PEAK1D DTHETA = 4. RTHETA = 1.5
REGION MERGING THRESHOLD (GLOBAL THETA) = 1.1



HIGH SENSITIVITY: ROAD SCENE 16. INTENSITY PLANE. NO ENHANCEMENT
WEAK SECTOR BOUNDARY MERGING ON TEST16I8P. 9-NOV-1985
SECTOR SIZE = 16. NONE ENHANCEMENT. DATA RANGE 0.0 TO 49.0
MIN-PEAK-DISTANCE 7 PERCENT. PEAK1D DTHETA = 3. RTHETA = 1.5
REGION MERGING THRESHOLD (GLOBAL THETA) = 0.6

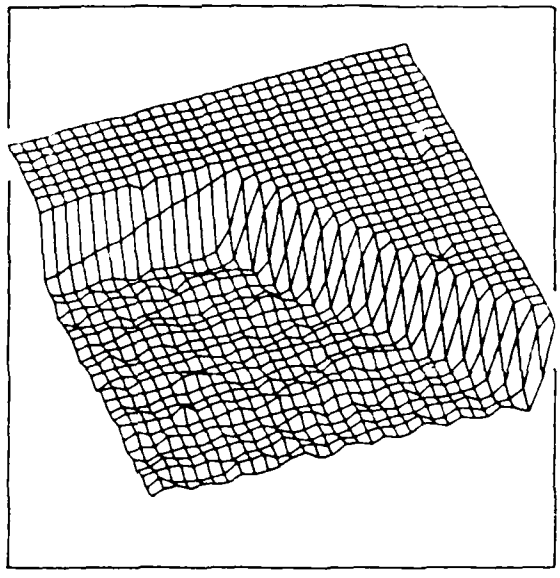


HIGH SENSITIVITY: ROAD SCENE 2 PIECE. RED COLOR PLANE
NORMAL SECTOR BOUNDARY MERGING ON TEST2RBP. 4-DEC-1985
SECTOR SIZE = 16. WEYMOUTH ENHANCEMENT. DATA RANGE 0.0 TO 80.0
MIN PEAK DISTANCE 7 PERCENT. PEAK1D DTHETA = 6. RTHETA = 1.5
REGION MERGING THRESHOLD (GLOBAL THETA) = 1.2

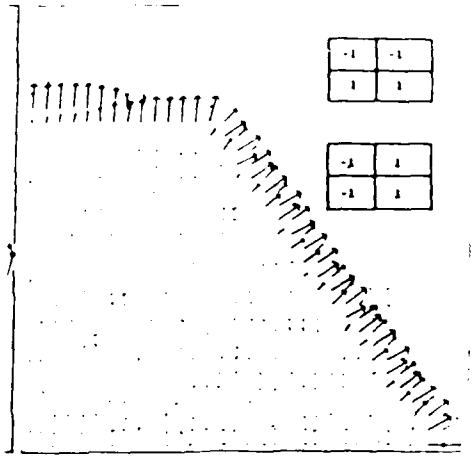


VERY HIGH SENSITIVITY: ROAD SCENE 10 PIECE. RED COLOR PLANE
NORMAL SECTOR BOUNDARY MERGING ON TEST10RBP. 9-DEC-1985
SECTOR SIZE = 16. WEYMOUTH ENHANCEMENT. DATA RANGE 1.0 TO 51.0
MIN-PEAK-DISTANCE 2 PERCENT. PEAK1D DTHETA = 1. RTHETA = 1.200000047
REGION MERGING THRESHOLD (GLOBAL THETA) = 1.2

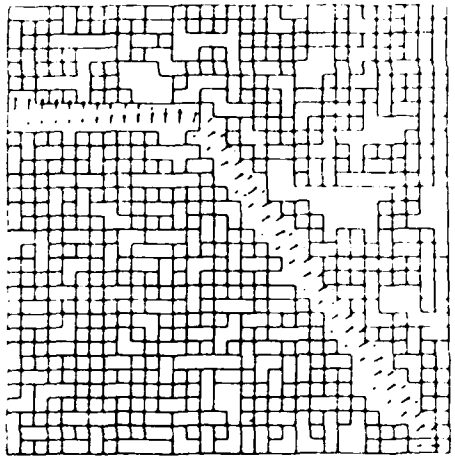
Figure 3.3



(a)



(b)



(c)

Figure 3.4

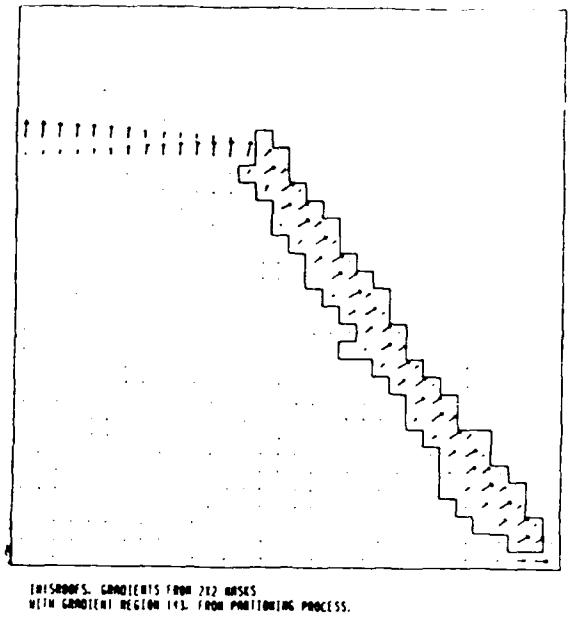
dient magnitude can be considered homogeneous region fragments, and adjacent elements could be grouped into homogeneous regions.

To make the fixed partition technique more sensitive to edges of any orientation, the algorithm actually uses two overlapping sets of partitions, with one set rotated a half-partition interval. Thus, if a 45 degree partition starting at 0 degrees is used, then a second set of 45 degree partitions starting at 22.5 is also used. The critical problem of this approach is merging the two representations in such a way that a single representation of each line is extracted from the two alternate line representations. The following scheme is used to select such regions for each line: first the lengths of the lines are determined for each line support region; then, since each pixel is a member of exactly two regions (one in each gradient segmentation), the pixel votes for the longest interpretation; finally the percentage of voting pixels within each line-support regions is the "support" of that region. Typically the regions selected are those that have support greater than 50%.

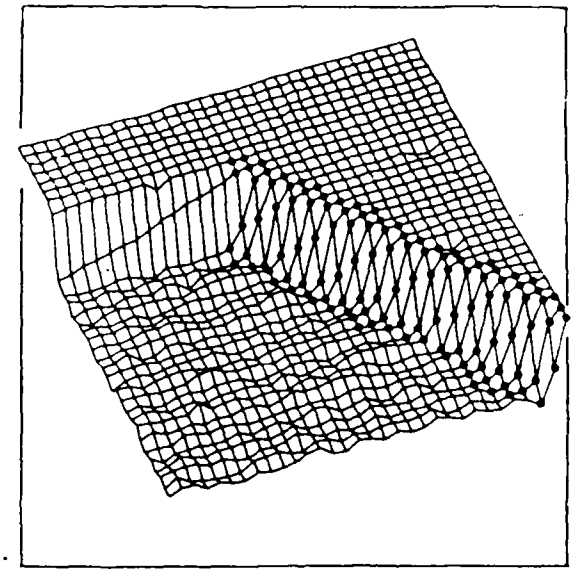
Initially, the underlying intensity surface of each line-support region will be assumed to have a meaningful straight line associated with it. In order to extract a representative straight line, a plane is fit to the intensity surface of the pixels in each edge-support region, using a least-squares method as in Haralick's planar surface patches of his slope-facet model [11]. An example region is depicted in figure 3.5(a) and as dots in the surface plot of figure 3.5(b). The pixels are weighted by local gradient magnitude in so that pixels in rapidly changing portions of the intensity surface would dominate the fit.

An obvious constraint on the orientation of the line is that it be perpendicular to the gradient of the fitted plane. A simple approach for locating the line along the projection of the gradient is to intersect the fitted plane with a horizontal plane representing the average intensity of the region weighted by local gradient magnitude as shown in Figure 3.5(c); the straight line resulting from the intersection of the two planes is shown in Figure 3.5(d).

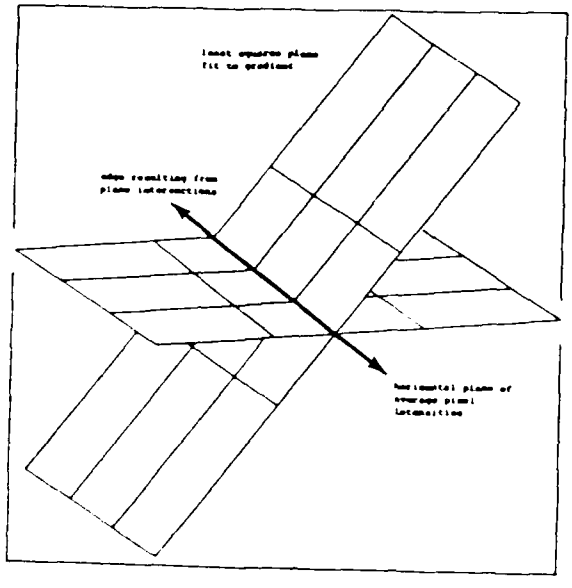
The line-support region and the planar fit of the associated intensity surface provides the basic



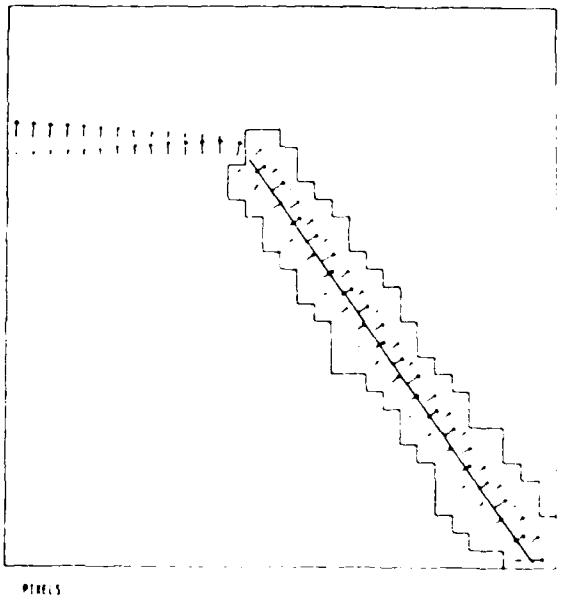
(a)



(b)



(c)



(d)

Figure 3.5

information necessary to quantify a variety of attributes beyond the basic orientation and position parameters. Length, width, contrast, and straightness can be easily extracted. Based upon these line attributes, the large set of lines can be filtered to extract a set with specific charateristics such as short texture-edges, or to select a "working set" of long lines at different levels of sensitivity.

The algorithm described in the preceding sections was applied to several full images shown in Figure 3.0. The algorithm is very robust and accurately extracts many low contrast long lines with overlapping partitions of 45 degrees, staggered by 22.5 degrees. Figure 3.6 demonstrates the performance of the algorithm.

Figure 3.6(a) shows the unfiltered output of the algorithm applied to a house image. Note that all of the small and low contrast edges are still present. We also show the result of filtering on the basis of gradient steepness (change in gray-levels per pixel) followed by a filtering on length that separates the edges into two disjoint sets, one corresponding to short texture edges (Figure 3.6(b)) and the other to longer lines related to the surface structure of objects in the image (Figure 3.6(c)). Figure 3.6(d) shows the result of filtering on length alone (length ≥ 5) for one of the other house scenes.

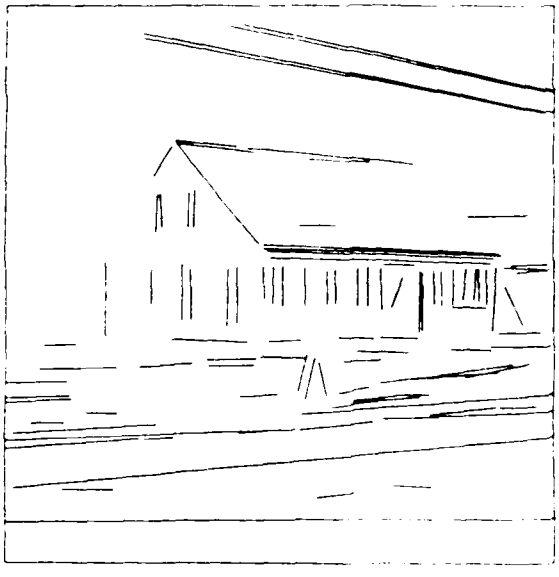
3.2.3 Rule-Based Region Merging

It became apparent during experimentation with the Nagin/Kohler histogram-based region segmentation algorithm that the best results were obtained by making the cluster selection very conservative and then merging most of the regions in the postprocessing stage; consequently, the merging process was extracted to form an independent system to be developed separately. The result of that development effort led to a rule-based merging algorithm, which is used to merge regions along the artificial sector boundaries the Nagin/Kohler algorithm, but which also is used with any algorithm which produces an overly fragmented segmentation.

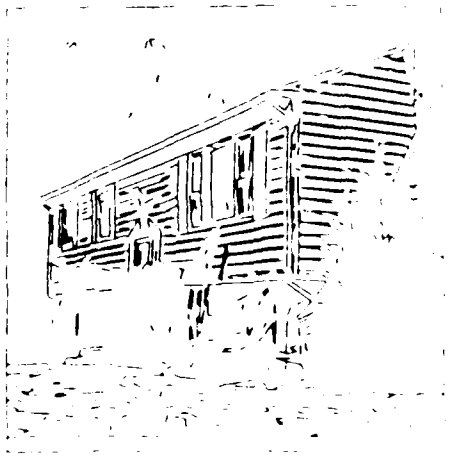
The rule-based region merging algorithm starts with an initial segmentation as input and selects pairs of adjacent regions which are candidates for merging into a single region. For example,

(a)

(b)



(c)



(d)

Figure 3.6

the initial segmentation can be produced by an extremely conservative region growing algorithm which produces a highly fragmented segmentation. Because the system only merges (as opposed to dividing) regions, the initial segmentation must contain all the region boundaries desired in the final result.

The merging phase of the algorithm is a locally iterative, and globally parallel, selection and testing of region pairs for merging. For each pair of regions the boundary between them is given a similarity score based on a similarity evaluation function. Then a search for all the local minimal boundaries is made. A boundary is considered locally minimal if and only if it is the lowest scoring boundary for both of its associated regions. Those minimal boundaries below the global threshold for region similarity are removed and the statistics for the newly created regions updated; removal of each boundary creates one new region from two old regions. The merging processes is applied iteratively and terminates when none of the remaining boundaries score below the global region similarity threshold.

The similarity function, which of course is the key to the effectiveness of the algorithm, takes into account both global and local information. The global features are extracted from the total populations of the pixels in the two regions, while the local features are extracted from the pixels in the neighborhood of the boundary between the two regions. The similarity function is defined as a set of rules, each of which measures the similarity of two regions on the basis of a single feature, such as color, intensity, common boundary length, boundary contrast or a combination of features. The results of many such rules are combined through a function which computes the overall similarity measure.

The combination function is the product of each of the similarity rules, each of which returns a value centered on one (implying no information about merging); values less than one indicate a vote for a merge (bounded by zero which guarantees a merge), and values greater that one indicate a vote against a merge. In practice the rules are never absolutely certain, and often return values

close to one.

Expectations of the characteristics of the images and goals regarding the desired segmentation can be encoded in the rules, and the nature of the segmentation produced can be controlled by the choice of rules and relative weights on the chosen rules. The philosophy of rule-based expert system methodology can be adapted here to allow modular contributions of features in a situation that is theoretically intractable (see discussion in introductory sections). By defining contributions from a feature set where the features can easily be varied for empirical evaluation or can be varied in situations where the goals of segmentation change, then we gain many of the advantages of expert system construction. Currently, the system includes seven rules that can be applied to a pair of regions to determine the desirability of a proposed merge:

1. Difference of the global means normalized by the sum of standard deviations of the candidate regions.
2. Region size - small regions are encouraged to merge with larger regions.
3. Degree of adjacency - regions connected by a relatively long boundary are encouraged to merge.
4. Similarity of the standard deviations.
5. Difference of local means.
6. Maximum allowable standard deviation of the merged region.
7. The degree to which the region could have been caused by mixed pixels during the digitization process (i.e. a narrow region between regions of locally lower and higher means).

Figure 3.7 shows the results of applying the merging algorithm to the output of a conservative region segmentation generated by the Nagin/Kohler algorithm described in Section 3.2.1. Figure 3.8 shows the final results for several house scenes.

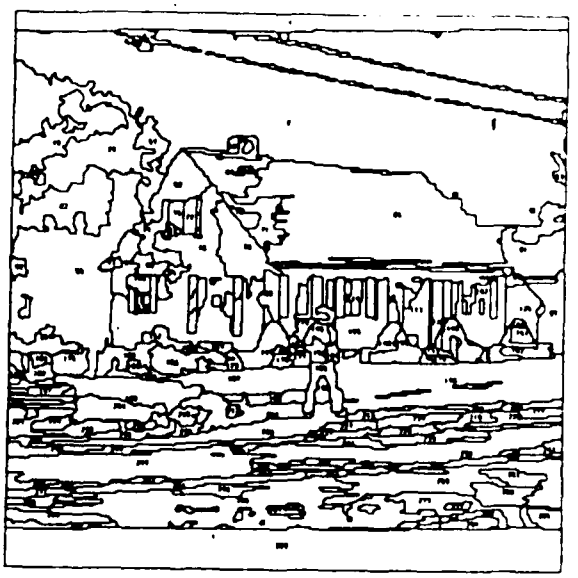

THREE COLOR UNION: TEST2U8P:WEY:WIG:NORMAL

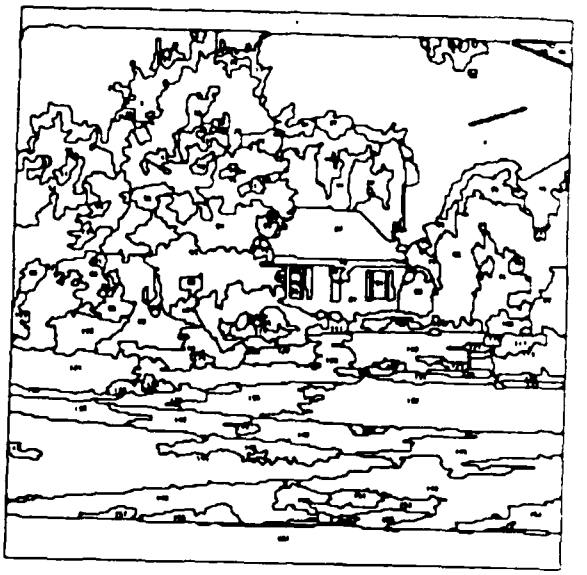
(a)


LOCAL HIST UNION RULE REMERGED

(b)

Figure 3.7





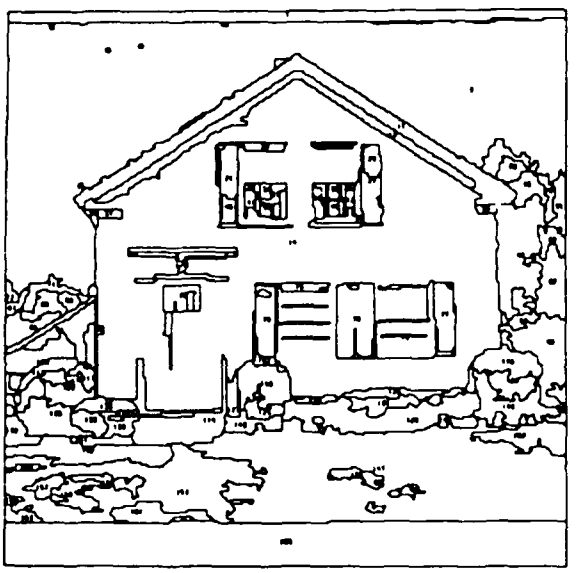


Figure 3.8

It should be noted that our approach differs significantly from the approach taken by Nasif [22], although both systems are rule based. Nasif's system is more general in that it allows both splitting and merging, and in that it represents both lines and regions. It is our position that it is not computationally reasonable to use a production rule paradigm as a full segmentation processs; Nasif's rules are of the form "If condition, Then action." Our rules are all applied simultaneously to any decision to merge a pair of regions, but require an external control structure to decide globally which merge they execute.

3.2.4 Low-Level Executive

The varying goals, features, and algorithms and their parameters makes the problem of delivering the best intermediate representation to the interpretation proceses an intractable problem. The "best" choices will almost certainly vary across images and vary across different locations in a single image. Consequently, one cannot expect to properly extract an optimal, good, or complete set of image events. Rather, one only hopes to deliver a usable representation.

We are in the process of developing the feedback loops from interpretation processes to a low-level executive which has knowledge about the set of low-level algorithms that we have described. The low-level executive will also be able to accept goal-oriented requests for re-analyzing the sensory data in particular locations in order to extract tokens with particular characteristics. If successful, this capability will change the analysis of the sensory data in a fundamental way.

4 Image Interpretation

The use of world knowledge, together with top down control, is beneficial and probably essential in domains where uncertain data and intermediate results containing errors cannot be avoided. Ambiguity and uncertainty in image interpretation tasks arise from many sources, including the inherent variation of objects in the natural world (e.g., the size, shape, color and texture of trees), the ambiguities arising from the perspective projection of the 3D world onto a 2D image plane, occlusion, changes in lighting, changes in season, image artifacts introduced by the digitization process, etc. Nevertheless, even with marginal bottom-up information, in familiar situations human observers can infer the presence and location of objects.

In the VISIONS system convergent evidence from multiple interpretation strategies is organized by top-down control mechanisms in the context of a partial interpretation. The extreme variations that occur across images can be compensated for somewhat by utilizing an adaptive strategy. This approach is based on the observation that the variation in the appearance of objects (region feature measures across images) is much greater than object variations within an image. The use of exemplar strategies using initial hypotheses and other top-down strategies results in the extension of partial interpretations from islands of reliability. Finally a verification phase can be applied where relations between object hypotheses are examined for consistency.

The interpretation task we are concerned with here is that of labelling an initial region segmentation of an image with object (and object part) labels when the image is known to be a member of a restricted class of scenes (e.g., suburban house scenes). The systems developed by Ohta [24] for understanding images of buildings in outdoor settings and by Nagao [20] for understanding aerial photographs bear some similarity to the techniques employed here. A review of these and other related work in image interpretation appears in [2].

4.1 A Knowledge Network and Representation Using Schemata

Description of scenes, at various levels of detail, are captured in a set of schema hierarchies [8]. A schema graph is an organizational structure defining an expected collection of objects, such as a house scene, the expected visual attributes associated with the objects in the schema (each of which can have an associated schema), and the expected relations among them. For example, a house (in a house scene hierarchy) has roof and house wall as sub-parts, and the house wall has windows, shutters, and doors as sub-parts. The knowledge network shown in Figure 4.1 is a portion of a schema hierarchy developed within the system. Each schema node (e.g. house, house wall, and roof) has both a structural description appropriate to the level of detail and methods of access to a set of hypothesis and verification strategies called interpretation strategies. For example, the sky-object schema (associated with the outdoor-scene schema) has access to the exemplar selection and extension strategy discussed below.

Interpretation rules relate image events to knowledge events by providing evidence for or against part/sub-part hypotheses. An interpretation strategy, associated with a schema node, specifies in procedural form how specific interpretation rules may be applied, and how combined results from multiple rules may be used to decide whether or not to "accept" (i.e., instantiate) an object hypothesis. An interpretation strategy thus represents both control local to the node and top-down control over the instantiation process.

Note that the goal is not to have these interpretation rules and strategies extract exactly the correct set of regions. Our philosophy is to allow incorrect, but reasonable, hypotheses to be made and to bring to bear other knowledge (such as various similarity measures and spatial constraints) to filter the incorrect hypotheses. An example of such error detection and correction in the interpretation process will be given later.

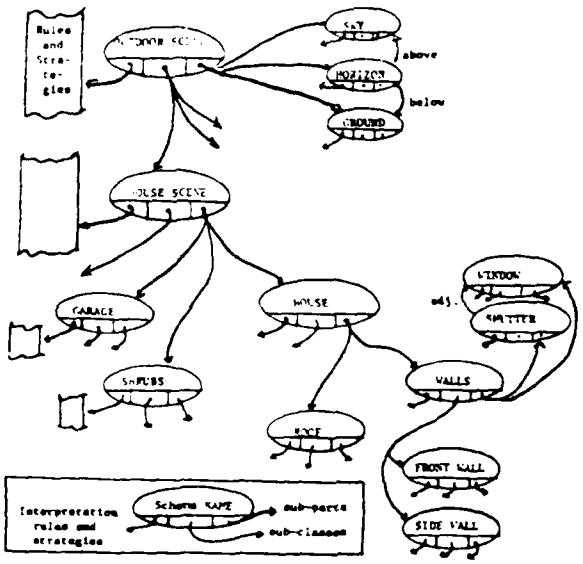


Figure 4.1 Abstract representation of a portion of the knowledge network used to produce the interpretation results. Knowledge about a class of scenes is represented by a hierarchy of schemata embeded in a network. Structural descriptions of each schema are organized by the component descriptions of subclass and subpart, and by spatial relations. Each schema node has acess to a set of interpretation rules which form hypotheses on the bases of image measurements, and interpretation strategies which describe how these hypotheses are combined with information in the network to form a consistent interpretation. Although every node in the network is considered to be a schema, only selected nodes in the figure show the full structure.

Figure 4.1

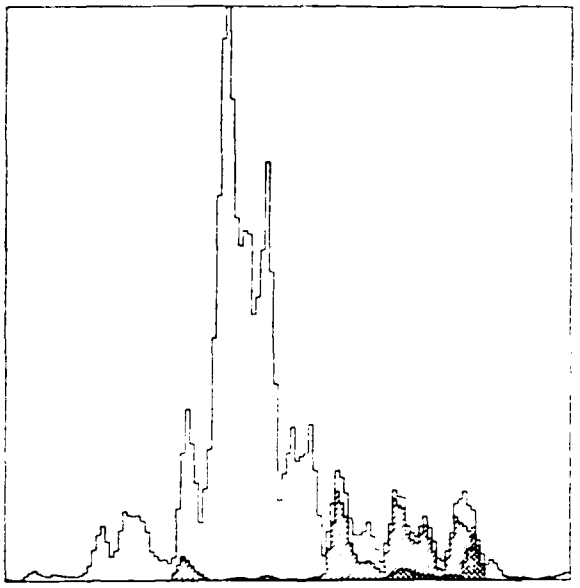
4.2 Rule Form for Object Hypotheses Under Uncertainty

We will illustrate the form of a simple interpretation rule based on using the expectation that grass is green. The feature used is average "excess green" for the region, obtained by computing the mean of 2G-R-B for all pixels in this region. Histograms of this feature are shown in Figure 4.2 comparing all regions to all known grass regions across 8 samples of color outdoor scenes. An abstract version is shown in Figure 4.3. The basic idea is to form a mapping from a measured value of the feature obtained from an image region, say f_I into a "vote" for the object on the basis of this single feature. One approach to defining this mapping is based on the notion of prototype vectors and the distance from a given measurement to the prototype, a well-known pattern classification technique which extends to N-dimensional feature space [8]. In our case rather than using this distance to "classify" objects in a pure Bayesian approach that is replete with difficulties, we translate it into a "vote".

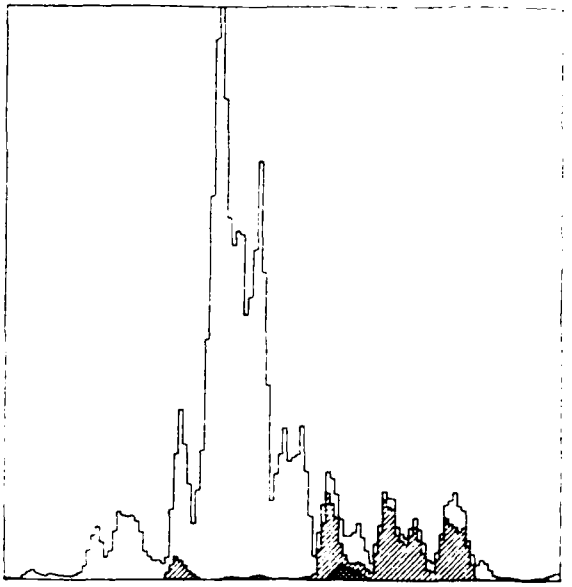
Let $d(f_P, f_I)$ be the distance between the prototype feature point f_P and the measured feature value f_I. The response R of the rule is then:

$$P(f_I) = \begin{cases} 1 & \text{if} \quad d(f_P f_I) \leq \Theta_1 \\ \frac{\Theta_2 - d(f_P f_I)}{\Theta_2 - \Theta_1} & \text{if } \Theta_1 < d(f_P f I) \leq \Theta_2 \\ 0 & \text{if } \Theta_2 < d(f_P f_I) \leq \Theta_3 \\ -\infty & \text{if } \Theta_3 < d(f P f I) \end{cases}$$

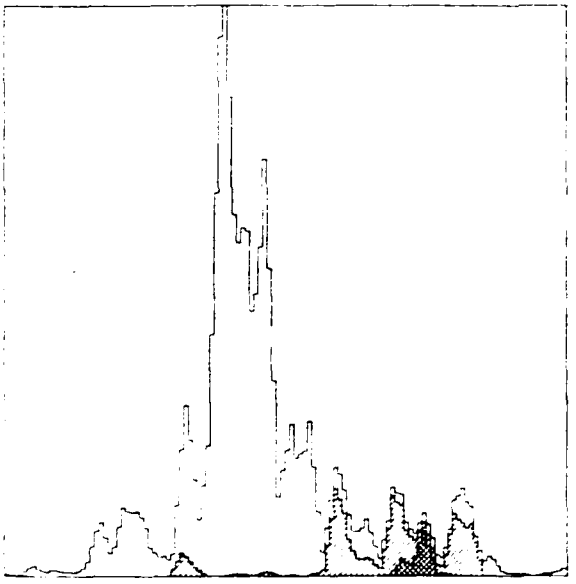
The thresholds Θ_1, Θ_2, and Θ_3 represent a gross mapping from the feature space to a score value that provides an interpretation of the distance measurements. Θ_3 allows strong negative votes if the measured feature value implies that the hypothesized object cannot be correct. For example, fairly negative values of the excess green feature imply a color which should veto the grass label. Thus, certain measurements can exclude object labels; this proves to be a very effective mechanism for filtering many spurious weak responses. Of course there is the danger of excluding the proper label due to a single feature value, even in the face of strong support from many other features. In the actual implementation of this rule form, $\Theta_1, \Theta_2, and \Theta_3$ are replaced with six values so that non-symmetric rules may be defined as shown in Figure 4.4. There are many possibilities



(a)



(b)



(c)

Figure 4.2

Image histograms of an "excess green" feature (2G-R-B) computed across eight sample images. The unshaded histogram represents the global distribution of the feature. The darkest cross hatched histogram is the distribution of this feature across regions known to be grass (from a hand labeling of the images) in one of three specific images. The intermediate cross hatching represents all known grass regions across the entire sample. Note the shifting (with respect to the full histogram) of the histograms for the individual images.

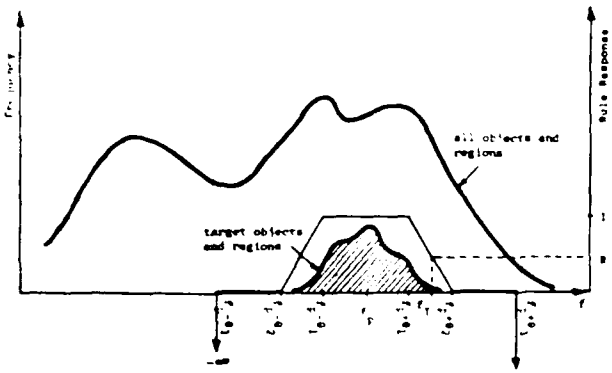


Figure 4.3 Structure of a simple rule for mapping an image feature measurement f_I into support for a label hypothesis on the basis of a prototype feature value obtained from the combined histograms of labeled regions across image samples. The object specific mapping is parameterized by four values, f_P, θ_1, θ_2, θ_3, and stored in the knowledge network. The use of six values will allow an asymmetric response function.

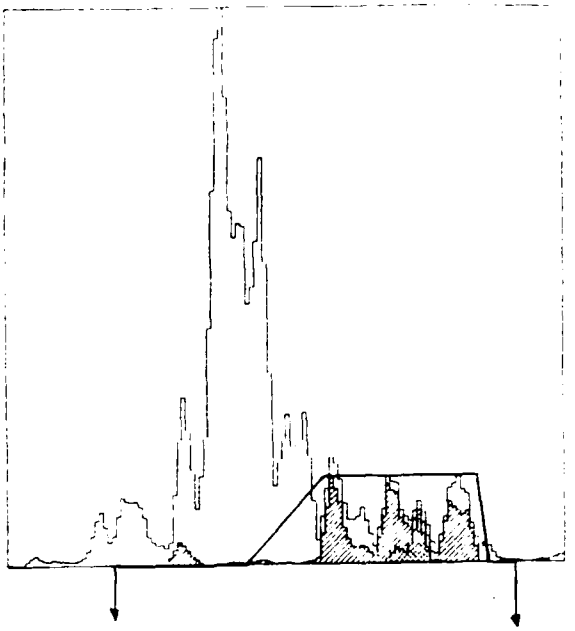
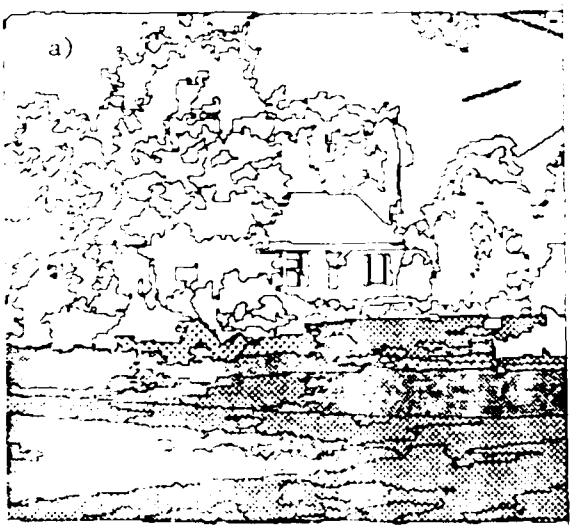
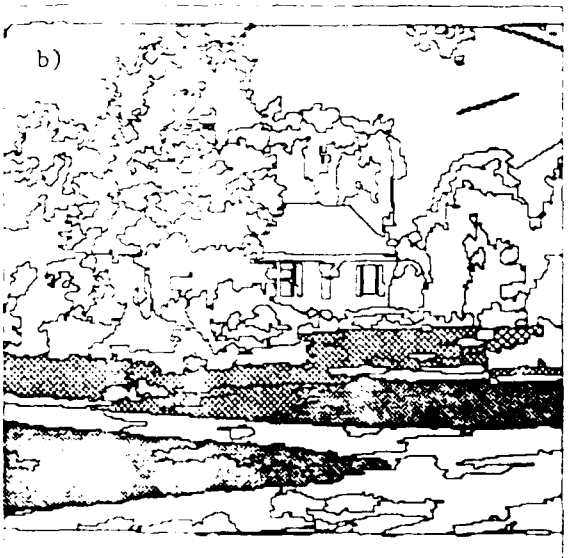


Figure 4.4 An example grass rule, showing an asymmetrical structure, superimposed on the histogram of Figure 2c.

Figure 4.5 The exemplar hypothesis rule is more selective than the corresponding general interpretation rule (based on a less selective rule form). Figure 4.5a shows the general grass interpretation rule, while Figure 4.5b shows the exemplar rule. Note that the general form of the rule results in more incorrect region hypothesis (which could be filtered by constraints from the knowledge network). Although the exemplar rule misses some grass regions, those found have high confidence.





for combining the individual feature responses into a score; here we have used a simple weighted average.

4.3 Exemplars and Islands of Reliability

The extreme variations that occur across images can be compensated for somewhat by utilizing an adaptive strategy. Variation in the appearance of objects (region feature measures across images) is much greater than object variations within an image (see Figure 4.2).

In the initial stages, there are few if any image hypotheses, and development of a partial interpretation must rely primarily on general knowledge of expected object characteristics in the image and not on the relationship to other hypotheses. The most reliable object hypotheses, derived from the interpretation rules, can be considered object "exemplars" and form basis of adaptation. One strategy extends the kernel interpretation by using the features of labelled exemplar regions including color, texture, shape, size, image location, and relative location to other objects. This is similar to the method in [19] where "characteristic regions" were used to guide hypothesis formation in the early stages of interpretation. The exemplar region (or set of regions) forms an image-specific prototype which can be used with a similarity measure to select and label other regions of the same identity. A verification phase can be applied where relations between object hypotheses are examined for consistency. Thus, the interpretation is extended through matching and processing of region characteristics as well as semantic inference.

Exemplar hypothesis rules differ from general hypothesis rules in that they are more conservative; they should minimize the number of false hypotheses at the risk of missing true target regions by narrowing their range of acceptable responses. If all regions are vetoed, secondary strategies are invoked; for example, the veto ranges can be relaxed, admitting less reliable exemplars. Figure 4.5 compares the results of the grass exemplar rule with the general grass hypothesis rule. The strategy can also be used to generate lists of hypotheses ordered by reliability.

The advantages of using object exemplars include:

1. an effective means for extending **reliable hypotheses** to regions which are more ambiguous; this is similar to the notion of **islands of reliability** [7].

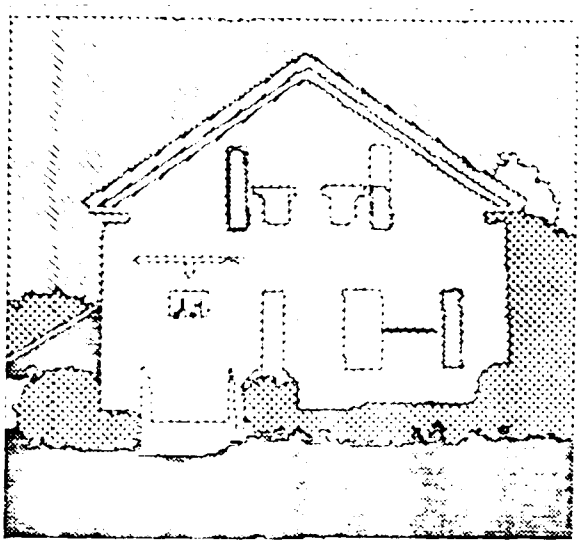
2. **knowledge-directed** technique for partially dealing with the unavoidable **region fragmentation** that occurs with any segmentation algorithm or low-level image transformation/grouping; regions that are "similar" to the exemplar can be both labelled and merged;

3. exemplars play a natural role in the implementation of an **hypothesize-and-verify control** strategy; hypotheses are formed based upon initial feature information and subsequently can be used in a verification process where the relationship between labelled regions provides consistency checks on the hypotheses and the evolving interpretation.

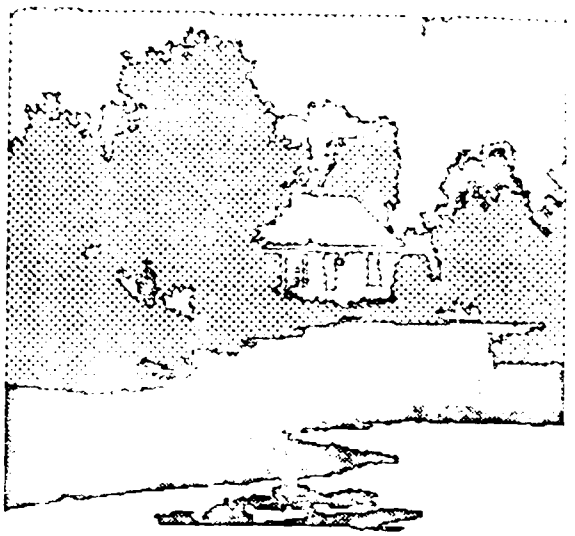
4.4 Results of Rule Based Image Interpretation

Experiments are being conducted on a set of eight "house scene" images. Thus far, we have been able to extract sky, grass, and foliage (trees and bushes) from five house images with reasonable effectiveness, and have been successful in identifying houses and their parts, including shutters (or windows), house wall and roof in three of these images. The interpretation strategies use many redundant features, each of which can very often be expected to be present. The premise is that many redundant features allow any single feature to be unreliable. Object hypothesis rules were employed as described in previous sections, while object verification rules requiring consistent relationships with other object labels are under current development. The final results shown in Figure 4.6 are an interpretation based on coarse segmentations. Further work on segmentation is being carried out, as is the refinement of the exemplar selection and matching rules.

An extremely important capability for an interpretation system is feedback to lower level processes for a variety of purposes. The interpretation processes should have focus-of-attention mechanisms for correction of segmentation errors, extraction of finer image detail, and verification of semantic hypotheses. An example of the effectiveness of semantically directed feedback to segmen-



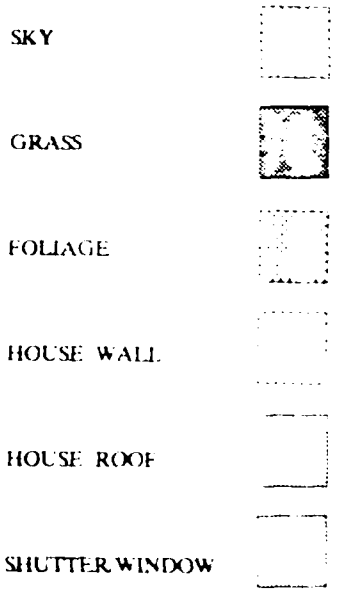
(a)



(b)

(c)

Figure . 6 Example interpretations for three of the house scene images. The labeling is



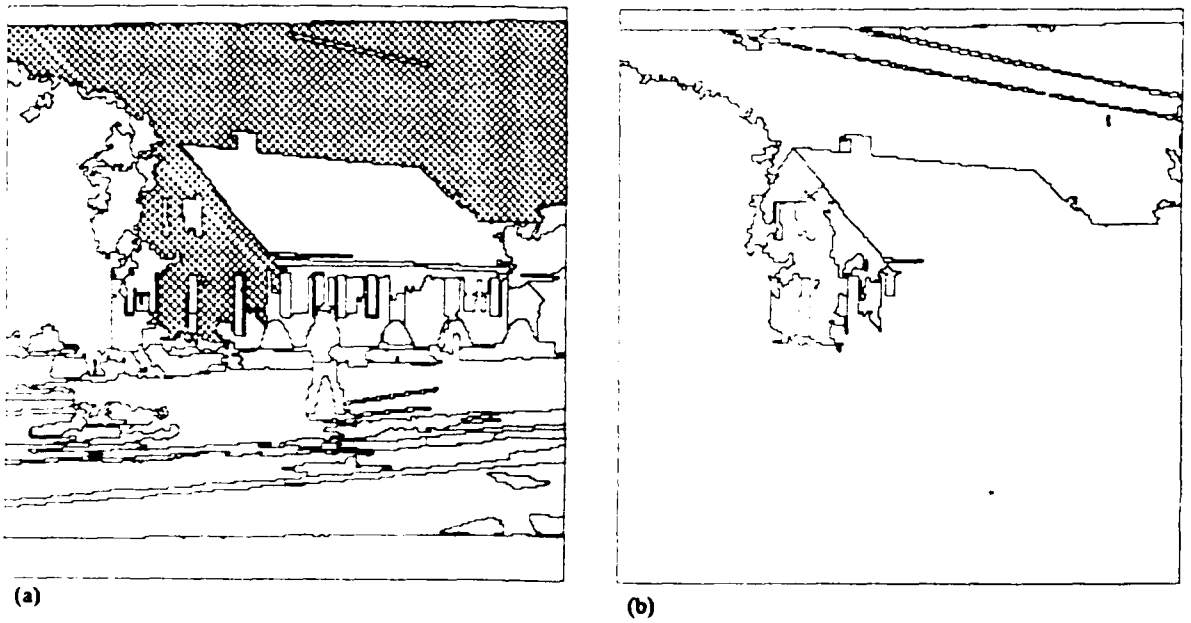


Figure 4.7 Resegmentation of house/sky region from Figure 6.c Figure 7a is the original segmentation showing the region to be resegmentated; 7b shows the regions resulting from the selective application to the segmentation process to the cross-hatched area in 7a.

tation processes is shown in Figure 4.7. There is a key missing boundary between the house wall and sky which leads to incorrect object hypotheses based upon local interpretation strategies. The region is hypothesized to be sky by the sky strategy, while application of the house wall strategy (using the roof and shutters as spatial constraints on the location of house wall) leads to a wall hypothesis.

There is evidence available that some form of error has occurred in this example: 1) conflicting labels are produced for the same region by local interpretation strategies; 2) the house wall label is associated with regions above the roof (note that while there are houses with a wall above a lower roof, the geometric consistency of the object shape is not satisfied in this example); and 3) the sky extends down close to the approximate horizon line in only a portion of the image (which is possible, but worthy of closer inspection).

In this case resegmentation of the sky-housewall region, with segmentation parameters set to extract finer detail, produces the results shown in Figure 4.7a. Subsequent remerging of similar regions produces a usable segmentation of this region as shown in 4.7b. It should be pointed out that in this image there is a discernable boundary between the sky and house wall. Initially, the segmentation parameters may be set so that the initial segmentation misses this boundary. This may occur because of computational requirements (fast, coarse segmentations) or as an explicit control strategy. However, once it is resegmented with an intent of overfragmentation, this boundary can be detected. Remerging based on region means and variances of a set of features allows much of the overfragmentation to be removed. Now, the same interpretation strategy used eariler produces quite acceptable results shown in Figure 4.8.

The current development of interpretation strategies involves the utilization of stored knowledge and a partial model (labelled regions) for hypothesis extension. In these strategies the knowledge network is examined for objects that can be inferred from identified objects, and for relations that would differentiate them. For example, the bush regions can be differentiated from other foliage

Figure 4.8 Final interpretation of the house scene in Figure 6c, after inserting resegmented house/sky regions and reinterpreting the image.

Figure 4.9 An example of the use of spatial relations to filter and extend region labeling. The geometric relations between house and shrub (in 9a) and between between roof and house front wall (in 9b) are used to refine region hypotheses from the interpretation shown in Figure 6c. Note that there are still ambiguities (the shrub label in the grass area, and the pants labeled as house wall) that require the use of other filters.

(a)

(b)

based on their spatial relations to the house, and front and side house walls can be differentiated using geometric knowledge of house structure (e.g., relations between roof and walls), as shown in Figure 4.9. In the full system, these rules would not work in isolation as shown here, and the errors made by this type of rule would be filtered by other constraints.

Future work is directed towards refinement of the segmentation algorithms, object hypothesis rules, object verification rules, and interpretation strategies. System development is aimed towards more robust methods of control: automatic schema and strategy selection, interpretation of images under more than one general class of schemata, and automatic focus of attention mechanisms and error-correcting strategies for resolving interpretation errors.

Terry Weymouth's forthcoming Ph.D. Thesis describes a schema-oriented system which interprets images of suburban house scenes. Interpretation results in a network with object descriptions as nodes and interobject relations as arcs. For each object, the object description includes the approximate location of the object in the scene and the location of the object in the image. The interpretation network is constructed by the cooperative activities of schemas in a schema network. Each schema describes a scene or object in two ways. A declarative structure describes the compositional and spatial relations, especially those of the types of parts, their potential spatial relations, and their possible appearance in an image. On the other hand, the schema represents, by references to special procedures called interpretation strategies, information on how and when to group image feature and create hypothesis of object existence.

The system is designed to apply a set of schema to an image, treating the activation of a schema as a separate invocation of the interpretation strategies. This invocation of the schema is called a schema instance. Schema instances can interact directly by the request of an interpretation goal, or indirectly by creating hypothesis and monitoring the interpretation network for the creation of hypothesis. Several schema instances can be active at the same time, each independently interpreting a portion of the image. Schema activation is controlled by both the interpretation strategies

(top down) and the presence or absence of data confirming the existence of the object (bottom up).

The system is being tested on seven images from four scenes. Preliminary results from this experiment show that this method can be used to interpret suburban house scenes. The resulting interpretations contain both three- dimensional descriptins of the objects in the scene (where appropriate) and the association between object and image. Hopefully, experiments will also show the potential for parallel activation of schema, with independent schema working on separate portions of the image.

5 Inferencing and the Inference Network

The construction of an image model is critically dependent on an ability to interpret the typically imperfect information provided by the various rules and interpretation strategies within the context of domain knowledge, system goals, and current hypotheses about the interpretation. An implicit assumption of our research is that the set of possible interpretations can be sufficiently constrained by some body of knowledge and inferences from the presence or absence of image features can be pooled correctly. A major factor contributing to this abiguity is the degree to which the knowledge sources provide conflicting evidence. It has been shown [8,25] that ambiguity arising from the lack of perfect information can be substantially reduced by obtaining partially redundant information from a variety of different sources. However, a major problem has been to develop mechanisms with some theoretical foundation that can take such unreliable and incomplete information and interpret it within the context of the available knowledge.

Some of the limitations of inferenceing using Bayesian probability models are overcome using the Dempster-Shafer formalism for evidential reasoning, in which an explicit representation of partial ignorance is provided [29]. The inferencing model allows "belief" or "confidence" in a proposition to be represented as a range within the [0,1] interval. The lower and upper bounds represent support and plausibility, respectively, of a proposition, while the width of the interval can be interpreted as ignorance.

Evidential information, extracted from the environment by modular sources of knowledge, enters these models in the form of probability "mass" distributions which are defined over sets of propositions common to both them and the model. These mass distributions are combined, relative to the possibilities embodied in the model, through Dempster's rule of combination [5]. The result is a new mass distribution representing the consensus of the information combined. This information is converted to the interval representation, and the model allows "inference" from those propositions it directly bears upon to those it indirectly bears upon (Figure 5.1). The apriori prob-

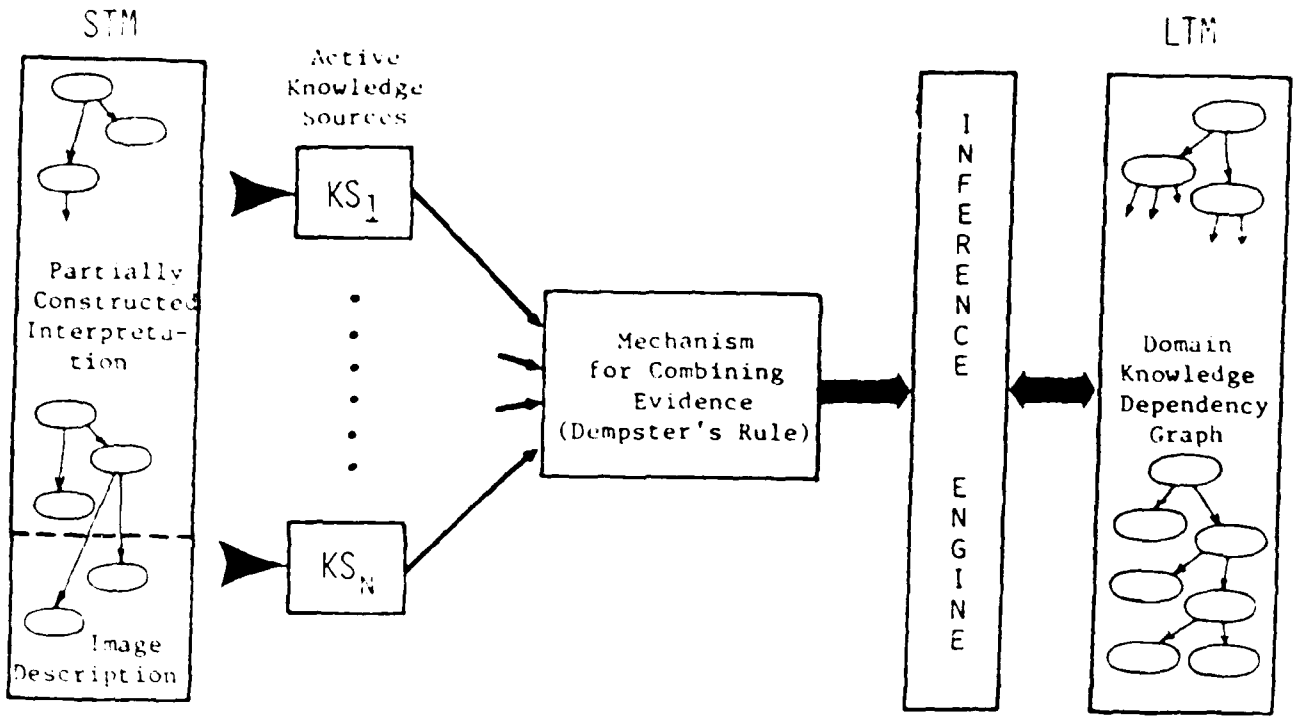


Figure 5.1 Architecture of the Inferencing Process.

Evidence for and against particular nodes (i.e., semantic concepts) in LTM is obtained via the application of modular knowledge sources to the partially completed interpretation in STM. Related evidence is combined and the results propagated through the domain knowledge represented as a hierarchically organized dependency graph in LTM. The resulting changes in the confidence levels attached to nodes in LTM may be used as the basis of a focus of attention mechanism and for system control.

abilities, frequently difficult or impossible to collect in artificial intelligence domains, but required by most other systems of inexact reasoning, are not needed. This form of evidential reasoning is more general than either a Boolean or Bayesian approach, yet it reduces to Boolean or Bayesian inferencing when the appropriate information is available.

Within the VISIONS system use of the DGMES model to reason about the environment involves five main steps:

1. Obtain evidence (for example, from the hypothesis rules) that tends to confirm or refute the truthfulness of hypotheses represented in a dependency graph (LTM).

2. Use Dempster's rule of combination [5,6] to pool the evidence into a form suitable for input to the inference engine.

3. Record the effect of the pooled evidence on hypotheses it bears directly upon.

4. Propagate the effect of the pooled evidence to the remaining hypotheses in the inference network by updating the confidence intervals associated with each semantic entity in LTM.

5. Determine the belief in each entity from the confidence interval.

There are many reasons why the evidential model is attractive. It separates the mechanisms for combining evidence from the mechanisms for making environmental inferenes, allowing us to experiment with ways to combine data that are independent of the representation of domain knowledge. The model also does not require perfect information; however, if it is available, it can be easily integrated with existing information. The model can perform both data and goal directed inferences over a single knowledge network. The theoretical foundation of the evidential model makes it easier to understand the relationship between the manipulation of environmental inormation and knowledge, and the performance of the system. Because of its formality it is easier to prove, if necessary, why the system performed the way it did, given some body of evidence and domain knowledge.

The inferencing system is implemented in GRASPER and has been applied to restricted cases of reasoning in the image interpretation domain [31]. We are currently examining ways in which the inference mechanism can be used to propagate the results of data-directed hypotheses through the long term knowledge structure leading torwards schema instantiation; at the same time we are exploring the use of the same mechanisms to propagate downward through the knowledge representation toward activation of bottom-up processes. Both of these represent focus-of-attention mechanisms which can be used by the schema-driven interpretation strategies to determine how the current partial interpretation can be most profitably extended. Thus, we see the inference engine as a plausible connection between data-directed and goal-directed hypothesis formation and instantiation. Wesley [32] is extending this approach to distributed control over the interpretation process; the view is that by describing the available system resources, control over the interpretation process can be achieved using a set of very general, domain independent goals.

5.1 Controlling the High Level Interpretation of Static Outdoor Natural Scenes: An Evidential Approach

Expert systems that are designed to interpret images are continually confronted with the problem of deciding how to allocate their limited resources in order to successfully complete their tasks. Being able to reach a decision requires, in part, having the capacity to "reason" about a set of alternative actions. Furthermore such reasoning must be done with "evidential" information - i.e., information that is to some degree uncertain, imprecise, and occasionally inaccurate. It has been argued that Shafer's theory of belief functions as a basis for carrying out such reasoning. However, to date, AI has not seen the development of any large scale system to test and either substantiate or refute such claims.

It is widely accepted that knowledge based system (KBSs) that operate in complex domains must "reason" from information that is to some degree uncertain, imprecise, and occasionally inaccurate, called "evidential" information [17]. Furthermore, each body of infomration is usually

generically distinct and is typically obtained from a variety of disparate sources, commonly called knowledge sources (KSs). The evidential information that KSs provide is derived, in part, from imperfect perceptions of their environment. And as such, can be viewed as partial evidence for or against the occurence of semantically meaningful events in some domain of interest. Given this reality, the degree to which a KBS successfully deals with real world problems depends, in part, on the technology it employs to reason from evidential information.

Wesley's thesis is concerned with the integration and evaluation of a technology that KBSs might use to complete two fundamental tasks. One task is to reason from evidential information in order to interpret (i.e., understand) the perceptions of its KSs. The second is to decide how to allocate its limited resources in order to successfuly complete the previous task. That is, we must anticipate that the complexity of the real world prohibits a KBS from understanding its perceptions in one fell swoop. Rather, "control-related" information must be obtained in order to help make decisions about the type, nature, quality, and quantity of the information that is required to interpret the perceptions of KSs. In the work reported here, the control-related information that a KBS must reason from is provided by control knowledge sources (CKSs). Similar to KSs the information that CKSs provide is derived, in part, from their perceptions of the state of the system and or the environment. As a consequence, such control related information is also evidential in nature. Thus, KBSs will be more successful at understanding their perceptions to the degree they employ technologies that are better suited than current techniques for reasoning from limited evidential information.

The concept of evidential reasoning (ER) was introduced be Lowrance and Garvey [17]. This evolving technology starts from the position that the acquisition of information by KMSs involves making imperfect perceptions of the environment. A KBS "understands" its world by perceiving it through a set of KSs. And because a system's perceptual machinery is not flawless, it follows that the information the KSs provide will be to some degree uncertain, imprecise, and occasionally

inaccurate - evidential in nature. This concept currently relies on the DS formalism as its model for representing and pooling KSs' beliefs that are based on their environmental perceptions. Thus the DS formalism is fundamental to ER-based models that KBSs might use to reason in their task domain.

There are two distinct reasoning processes that must be completed in this concept. One is to take a single body of evidence and propagate its effect from those propositions the evidence bears directly upon to those it indirectly bears upon. This allows inferences to be drawn about those propositions not directly affected by the evidence. This process is typically carried out by what is commonly called an inference engine. The other process, one that pools multiple bodies of evidence into a single body of evidence that represents a consensus opinion, is accomplished by Dempster's rule.

A flow diagram of the evidential-based high-level computer vision system (EHCVIS) is being explored by Wesley in Figure 5.2. As indicated by the figure, EHCVIS can be described in four phases. The task of the first phase is to use the specifications of goals to help complete two subtasks. Examples of goals the system might try to reach are finding a house, locating the ground plane, or obtaining additional information to help resolve some ambiguity the system might have about the identity of objects in a region of interest. The first subtask is to use goal specifications to generate a set of alternative actions the system might pursue in order to reach that goal. The second subtask is to use goal specifications to select control strategies that will be used to help decide which alternative action is more appropriate to pursue.

The second phase can be summarized in several steps:

1. with the alternative actions and control strategies that were selected in the previous phase, dynamically build the control knowledge that will be brought to bear on the problem of deciding which alternative to pursue;

2. implement these control strategies, in part, by obtaining control related information from

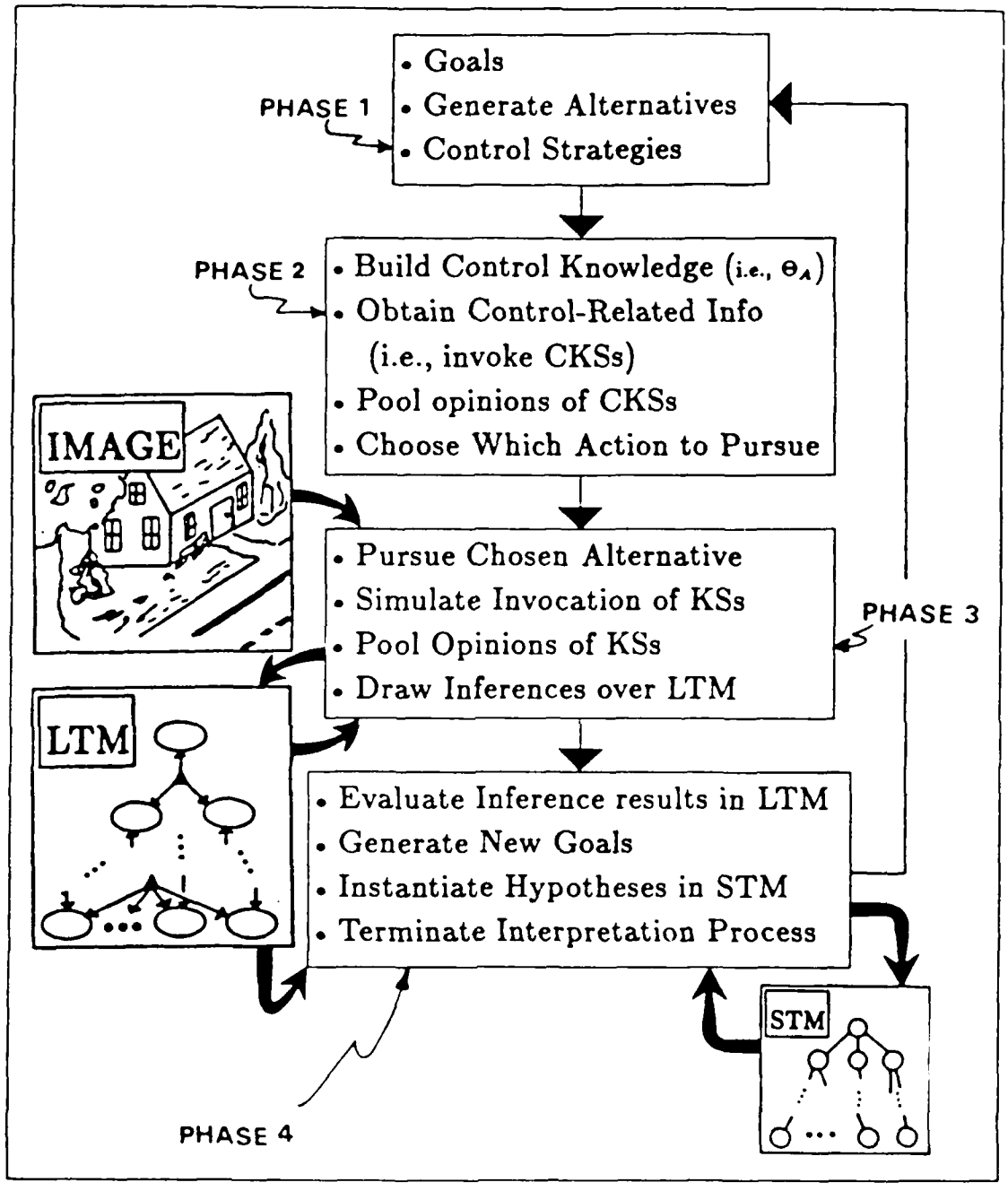


Figure 5.2

A SYSTEM FLOW DIAGRAM OF EHCVIS

independent control knowledge sources (CKSs); and

3. pool these beliefs using Dempster's rule and then use an inference engine to take the result of Dempster's rule and infer which action is the best to pursue.

Note that in the third step of the second phase (see Figure 5.2) beliefs are pooled and inferences are drawn over the control knowledge frame. This is to indicate that the DS technology is used by the proposed system to reason about its actions.

In the third phase, the system takes the action suggested by the second phase. A typical action might be to task a subset of available KSs to make some observation about a particular subset of regions in an image, then express some beliefs about their perceptions. After the KSs have done this, Dempster's rule is used to pool their beliefs and then inferences are drawn over LTM to infer which propositions (i.e., label hypotheses) in LTM should be associated with the region under examination.

In the last phase, the results of the inferences drawn over LTM are evaluated. Based on this evaluation the system might decide that a new goal should be satisfied and return to the goal generation phase. Or that the interpretation process should be terminated. Or that the system should "instantiate" (i.e., record in a dynamic representation called short term memory, STM) its belief that a subset of the label hypotheses in LTM should be associated with a subset of the regions in an image, and then set new goals to be satisfied.

There are several objectives of the research and experiments proposed by Wesley:

1. to demonstrate that certain types of incompletenesses in LTM can be detected when certain "evidential measures" and verification procedures are employed;

2. to demonstrate that the system tends to degrade smoothly as the quality of the information it must reason from becomes less certain, precise, and accurate, and;

3. to demonstrate that a system's performance is improved (i.e., fewer resources are used, better

interpretations, or a combination of the above) as more evidential-based control strategies are used.

A parameterized version of the EHCVIS based on the current VISIONS system is being built. When complete, it will be tested on a large set of reasonably complex outdoor image (primarily house scenes).

6 REFERENCES

REFERENCES

[1] R. Belknap, E. Riseman, and A. Hanson, "The Information Fusion Problem and Rule-Based Hypotheses Applied To Complex Aggregations of Image Events," *Proc. DARPA IU Workshop*, Miami Beach, FL, December 1985.

[2] T. Binford, "Survey of Model-Based Image Analysis Systems," *International Journal of Robotics Research*, **1**, 1982, pp. 18-64.

[3] M. Brady, "Computational Approaches to Image Understanding," *Computing Surveys*, **14**, March 1982, pp. 3-71.

[4] J.B. Burns, A. Hanson, and E. Riseman, "Extracting Linear Features," *Proc. 7th ICPR*, Montreal, 1984. Also COINS Technical Report 84-29, August 1984.

[5] A.P. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals of Mathematical Statistics*, 38, 1967, pp. 325-339.

[6] A.P. Dempster, "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society*, Series B, Vol. 30, 1968, pp. 205-247.

[7] L. Erman, et al., "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty," *Computing Surveys*, **12(2)**, June 1980, pp. 213-253.

[8] A. Hanson, and E. Riseman, "VISIONS: A Computer System for Interpreting Scenes, Computer Vision Systems, (A. Hanson and E. Riseman, eds.) pp. 303-333, Academic Press, 1978.

[9] A. Hanson and E. Riseman, "Segmentation of Natural Scenes," *Computer Vision Systems*, (A. Hanson and E. Riseman, Eds.), Academic Press 1978, pp. 129-163.

[10] A. Hanson and E. Riseman, "A Summary of Image Understanding Research at the University of Massachusetts," COINS Technical Report 83-35, October 1983, University of Massachusetts at Amherst.

[11] R.M. Haralick, "Ridges and Valleys on Digital Images," *Computer Vision Graphics and Image Processing*, 22(1), 1983, pp. 28-39.

[12] B.K.P. Horn, "Obtaining Shape from Shading Information," in *The Psychology of Computer Vision*, P.H. Winston (Ed.), McGraw-Hill, New York, 1975.

[13] B.K.P. Horn, "Understanding Image Intensitites," *Artificial Intelligence*, Vol. 8, 1977, pp. 201-231.

[14] C. Kohl, Ph.D. Dissertation, COINS, University of Massachusetts at Amherst, in preparation.

[15] R.R. Kohler, "Integrating Non-Semantic Knowledge into Image Segmentation Processes," COINS Technical Report 84-04, University of Massachusetts at Amherst, March 1984.

[16] V.R. Lesser, R.D. Fennell, L. D. Erman, and D.R. Reddy, "Organiztion of the Hearsay-II Speech Understanding System," *IEEE Trans. on ASSP 23*, 1975, pp. 11-23.

[17] J. Lowrance, "Dependency Graph Models of Evidential Support," Ph.D. Thesis, University of Massachusetts, Amherst, 1982; also COINS Technical Report No. 82-26.

[18] D. Marr, *VISION*, W.H. Freeman and Company, San Francisco, 1982.

[19] M. Nagao and T. Matsuyama, "Edge Preserving Smoothing", *Proc. of the Fourth International Joint Conference on Pattern Recognition*, pp. 518-520, November 1978.

[20] M. Nagao and T. Matsuyama, "A Structural Analysis of Complex Aerial Photographs," Plenum Press, New York, 1980.

[21] P.A. Nagin, "Studies in Image Segmentation Algorithms Based on Histogram Clustering and Relaxation," COINS Technical Report 79-15, University of Massachusetts at Amherst, September 1979.

[22] A.M. Nasif and M.D. Levine, "Low Level Segmentation: An Expert System," Technical Report 83-4, April 1983, Electrical Engineering, McGill University.

[23] R. Ohlander, K. Price, and D.R. Reddy, "Picture Segmentation Using a Recursive Region Splitting Method," *CGIP 8,3*, December 1979.

[24] Y. Ohta, "A Region-Oriented Image-Analysis System by Computer," Ph.D. Thesis, Computer Information Science Department, Kyoto University, Kyoto, Japan, 1980.

[25] C.C. Parma, A.R. Hanson and E.M. Riseman, "Experiments in Schema-Driven Interpretation of a Natural Scence," COINS Technical Report 80-10, April 1980, University of Massachusetts at Amherst.

[26] K.E. Price, "Changing Etection and Analysis in Multispectral Images," Ph.D. Thesis, Carnegie-Mellon University, Pittsburgh, PA, December 1976.

[27] G. Reynolds, D. Strahman, N. Lehrer, "Converting Feature Values to Evidence," *Proc. DARPA IU Workshop*, Miami Beach, FL, 1985.

[28] E.M. Riseman and A.R. Hanson, "A Methodology for the Development of General Knowledge-Based Vision Systems," *IEEE Proc. of the Workshop on Computer Vision: Representation and Control*, 1984, pp. 159-170.

[29] G. Shafer, "A Mathematical Theory of Evidence," Princeton University Press, 1976.

[30] J.M. Tenenbaum, H.G. Barrow, "Recovering intrinsic scene characteristics from images," in *computer Vision Systems*, (A. Hanson and E. Riseman Eds.), New York: Academic Press, 1978, pp. 3-26.

[31] L. Wesley and A. Hanson, "The Use of Evidential-Based Model for Representing Knowledge and Reasoning about Images in the VISIONS System," *Proc. Workshop on Computer Vision*, Rindge, NH, August 23-25, 1982.

[32] L. Wesley, "Reasoning About Control: An Evidential Approach", SRI Tech. Memo, to appear.

[33] T.E. Weymouth, J.S. Griffith, A.R. Hanson and E.M. Riseman, "Rule Based Strategies for Image Interpretation," *Proc. of AAAI-83*, August 1983, pp. 429-432, Washington D.C. A longer version of this paper appears in *Proc. of the DARPA Image Understanding Workshop*, June 1983, pp. 193-202, Arlington, VA.

[34] T. Williams, "Computer Interpretation of a Dynamic Image from a Moving Vehicle," Ph.D. Thesis and COINS Technical Report 81-22, University of Massachusetts at Amherst, May 1981.

7 APPENDIX 1

PUBLICATIONS SUPPORTED BY AFOSR CONTRACT F49620-83-C-0099

BUR84 Burns, J.B., Hanson, A.R. and Riseman, E.M., "Extracting Straight Lines," *Proc. of the IEEE Seventh International Conference on Pattern Recognition*, Montreal, Canada, July 30 - August 2, 1984, pp. 482-485; also COINS Technical Report 84-29, December 1984.

HAN83 Hanson, A.R. and Riseman, E.M., "A Summary of Image Understanding Research at the University of Massachusetts", COINS Technical Report 83-35, October 1983.

REY84 Reynolds, G., Irwin, N., Hanson, A. and Riseman, E., "Hierarchical Knowledge-Directed Object Extraction Using a Combined Region and Line Representation", *Proc. of the IEEE 1984 Workshop on Computer Vision Representation and Control*, Annapolis, Maryland, April 30 - May 2, 1984, pp. 238-247.

RIS84 Riseman, E.M. and Hanson, A.R., "A Methodology for the Development of General Knoweldge-Based Vision Systems", *Proc. of the IEEE Workshop on Principles of Knowledge-Based Systems*, Denver, Colorado, December 1984.

SHA83 Shaw, G.B., "Some Remarks on the Use of Color in Machine Vision", COINS Technical Report 83-31, September 1983.

WES85 Wesley, P., "Controlling the Highlevel Interpretation of Static Outdoor Natural Scenes: An Evidential Approach," Ph.D. Dissertation, COINS Department, University of Massachusetts, 1985.

WEY83 Weymouth, T.E., Griffith, J.S., Hanson, A.R. and Riseman, E.M., "Rule-Based Strategies for Image Interpretation", *AAAI 1983 Conference*, Washington, D.C., August 1983, pp. 429-432.

END

5-87

DTIC